제2권

데이터구축 안내서

2022. 2

과학기술정보통신부

NIA 한국지능정보사회진흥원

목 차

Ⅰ
크라우드소싱 기반의 작업자 관리

제1장 작업자 선발

1 크라우드워커의 특징

- 크라우드소싱 기반의 인공지능 학습용 데이터 구축 프로젝트는 불특정 다수의 일반인들이 참여하기 때문에 선발된 작업자들은 서로 다른 작업 숙련도와 참여 동기 등을 가지고 있다. 인공지능 학습용 데이터 구축사업에 적합한 작업자를 선발하여 투입하는 것이 데이터 품질을 확보하는데 매우 중요한 요소 중 하나이다.
- 외국어 번역이나 MRI 사진 바운딩박스 등과 같이 구축해야 하는 인공지능 학습용 데이터의 유형이나 해당 산업 분야에 따라서, 크라우드워커가 일반인이 아닌 특정 분야의 전문 역량이나 자격을 갖추어야 하는 경우도 있다.
- 불특정 다수의 일반인들을 모집하여 진행하는 경우, 데이터의 특성과 산업 분야에 적합한 인력을 선발해야 한다. 프로젝트 작업의 성격과 작업자의 역량 및 특성 간의 관계를 긴밀하게 파악해 적절하게 작업을 부여하는 것은 전체 프로젝트의 성공의 중요한 요인이다.
- 작업자의 성별, 나이 등 인구학적 속성과 개별 성격에 따라 작업의 품질이 달라지거나, 한 작업자의 개별 특성을 잘 고려하지 못하면 작업 완료된 데이터의 품질이 저하되는 경우도 많다. 따라서 역량 및 특성을 파악하여 목표로 하는 인공지능 학습용 데이터 구축에 적합한 작업자를 선발하고 그에 적절한 작업을 부여하는 것이 데이터 품질에 큰 영향을 미친다.
- 작업자 선발을 위해서는 목표로 하는 인공지능 학습용 데이터를 구축하기 위한 작업을 세분화하여 잘 정의하고, 적합한 작업자의 조건을 파악해야 하며, 다양한 방법을 통하여 적절한 인력을 선발해야 한다.
- 또한, 선발된 작업자와 검수자의 작업 역량을 개선하고 작업의 신뢰성을 확보하며 데이터 품질을 높이기 위해서는 작업자와 검수자를 대상으로 적절한 교육을 진행해야 한다. 학습용 데이터 구축사업 경험을 가지고 있다고 해도 작업 절차에 미숙하거나 명확한 작업

목표와 가이드를 숙지하지 못하여 데이터 품질이 좋지 않은 경우가 많으며, 검수자의 경우 검수에 대한 명확한 기준과 객관성을 확보해야 하므로 검수자를 위한 교육도 필요하다.

2 작업 설계 절차

- 인공지능 학습용 데이터 구축 생애주기에 따라서 작업자의 작업이 정의되어야 한다. 데이터 구축 생애주기별 작업 설계를 위해 데이터 구축을 위한 파일럿 프로젝트를 진행하여 작업 방법 설계, 작업 단위 설계, 작업 검사기준 정의, 작업 단가 설계 등을 해야 한다.
- 일반적인 작업 설계 절차는 '작업 방법을 설계'하고 '작업 단위를 정의'하고 이를 검수하기 위한 '작업 기준을 정의'하여야 한다. 이때 작업 완료가 되는 통과/반려/불가에 대한 기준을 수립하여야 한다.
- 작업 설계와 검수 설계를 마치면 작업 가이드를 작성하여 설계된 작업과 검수가 제대로 가능한지, 얼마나 가능한지 등을 측정할 수 있도록 파일럿 프로젝트를 수행한다. 이때 파일럿 프로젝트는 실제로 원시데이터를 수집하여 정제를 통해 원천데이터를 확보해야 하며, 원천데이터를 대상으로 어노테이션 도구를 사용하여 실제 데이터 라벨링을 하고, 라벨링데이터를 인공지능 학습 모델을 통하여 유효성 목표를 달성 가능한지 확인하는 1-Cycle 검사로 진행되어야 한다.
- 수집/정제/가공한 학습용 데이터가 인공지능 학습 모델에 적합한 데이터인지 1-Cycle 검증을 하고, 그 결과를 바탕으로 작업을 설계하여 작업 수행에 필요한 인력을 정의하고 작업 단가를 설계하여 크라우드워커를 선발하고 투입하여 실제 인공지능 학습용 데이터 구축 작업을 진행한다.
- 작업 설계가 적절한 지에 대해서는 반드시 파일럿 프로젝트를 통하여 검증이 필요하다. 실제 설계한 작업 방법이 적절하지 못하거나 난이도가 높아서 실제 크라우드워커를 투입했을 때 예상과 다르게 작업 속도가 느리거나 작업 품질이 떨어지는 경우가 발생할 수 있다. 이를 예방하기 위해서는 인공지능 학습용 데이터 구축 생애주기에 따른 적절한 작업 방법이 잘 설계되었는지 파일럿 프로젝트를 통하여 검증해야 하며, 이때 작업과 검수가 가능한 건 수 등을 추정하여 투입해야 하는 인원과 작업 단가를 설계한다. 작업 단가의 경우 작업 건수를 기준으로 설계하되 시간당 최저 임금보다 높은 수입이 확보될 수 있도록 설계되어야 크라우드워커들이 낮은 작업 단가로 인해 이탈하는 문제를 예방할 수 있다.

2.1 작업 설계 절차

- 인공지능 학습용 데이터 구축을 위한 작업 설계 절차는 구축 생애 주기(수집-정제-가공-검수-활용)에 따라서 진행된다.

[그림 Ⅰ-1] 작업 설계 절차

2.2 단계별 작업 설계

- 수집 : 작업자가 원시데이터를 확보하기 위하여 해야 하는 작업을 정의한다. 예를 들어 차량의 외관을 촬영하여 원시데이터를 확보하는 경우 차종, 연식, 정면, 측면, 촬영 시각도 등의 차량 외관 촬영 외에 부가적인 작업을 잘 정의하여 수집한 데이터를 정제하고 가공할 때 작업 불가 대상이 되지 않도록 정확한 작업 방법을 정의해야 한다.
- 정제 : 원시데이터를 포맷 통일, 중복 제거, 비식별화 등 일련의 전처리 과정을 통해 '원천데이터'를 확보하는 작업을 정의한다. 중복 데이터를 필터링할 때 중복 데이터의 기준을 제시하고 비식별화의 경우 관련 법규에 따른 비식별화 기준을 확인하여 제시하고 작업 불가 데이터에 대한 필터링 기준 등을 명확하게 제공하여야 한다.
- 가공 : 원천데이터를 인공지능 모델 학습을 위해 사용 목적에 따라 라벨링데이터를 구축하는 작업을 정의한다. 라벨링데이터는 어떤 데이터를 어떻게 가공하는지에 작업 단위를 정의하고 이를 어노테이션 도구를 사용해서 어떻게 가공하는지 도구에 대한 사용 방법과 함께 작업 방법을 설계해야 한다. 이때 검수의 기준이 되는 통과, 반려, 불가에 대한 기준이 함께 정의되어야 한다.
- 검수 : 작업자의 작업 결과물에 대해서 통과되는 기준, 반려되는 기준, 작업 불가 데이터에 대한 명확한 기준을 수립해야 한다. 이는 수집, 정제, 가공 단계의 작업 방법에 모두 영향을 미치므로 초기 단계에서 여러 차례 파일럿 프로젝트를 통하여 적절한 기준을 설계해야 한다.

3 선발 기준

- 크라우드워커 모집을 위해서는 데이터 구축과 관련된 작업자와 검수자에 대한 선발 기준을 수립하여야 한다. 특히 크라우드소싱 작업에서 작업자의 개별 특성 및 숙련도는 작업 품질에 큰 영향을 미치기 때문에, 프로젝트 관리자는 특정 작업에 적합한 조건을 가진 작업자를 선별해야 한다.

3.1 작업 단위 선발

- 선발 기준은 작업 단위에 맞추어 적절한 인력을 선발해야 한다. 작업 단위라는 것은 가장 최소 단위로 작업 방법 및 기준에 따라 작업을 정의해야 한다. 예를 들어 도로 주행 이미지에서 차량, 차선 등의 클래스 개수나 차량 개수 등의 오브젝트의 수 또는 차량, 차종 등 바운딩 대상이 되는 단위 등이 작업 단위의 기준이 될 수 있다. 작업 단위는 향후 작업자들의 단가를 산정하거나 생산성 측정 단위 등의 기준이 될 수 있기 때문에 작업 방법과 작업 대상을 고려하여 정의해야 한다.

3.2 검수자 우선 선발

- 작업자들에 앞서 작업 방식 및 검사기준에 대한 검증을 담당하는 인원으로 향후 품질 보증을 담당하는 핵심 인력인 검수자를 우선 선발해야 한다. 검수자는 기본적으로 검사 대상이 되는 작업에 숙련된 작업자여야 하므로, 선발 시 작업에 투입하여 충분히 작업 방법이나 기준에 익숙해질 수 있도록 해야 한다. 특히, 크라우드워커 중에서 인공지능 학습용 데이터 구축사업을 경험한 인력 중에서 선발하는 것이 바람직하다.
- 검수자는 작업에 대한 검사를 수행하면서 통과나 반려를 통한 재작업 여부를 판단하기 때문에, 작업 가이드나 작업 환경이나 목표에 대한 명확한 이해와 숙지가 이루어져야 한다. 또한 작업 결과에 대한 재작업 요청 시, 오류 항목에 대한 지적과 시정 방향을 제시할 수 있어야 하므로 다른 작업자와 소통과 협업이 가능한 인력으로 선발해야 한다.

3.3 작업 관련 인력 특성

- 작업자의 개별 특성과 작업 숙련도는 작업의 양과 질에 큰 영향을 미치기 때문에, 데이터 품질 향상을 위해서 필수적으로 고려해야 하는 요소이다. 작업자의 성별, 나이 등 인구학

적 속성과 성격에 따라 작업물의 난이도 별로 작업 품질이 상이할 수 있다. 작업자의 개별 특성 및 숙련도를 적절히 고려하지 못했을 때, 데이터의 품질이 저하됨을 알 수 있다. 수행해야 하는 데이터 작업과 관련하여 산업분야 지식이나 성별, 학력, 직업 등의 작업과 연관성 있는 인원으로 선발해야 한다.

3.4 우대 조건 고려

- 작업의 특성을 고려하여 작업자 및 검수자의 우대 조건을 정의하여 인력 모집에 활용해야 한다. 작업 대상이 차량의 부품별 외관 사진을 촬영하는 일이라면, 차량에 대한 기본적인 지식을 가지고 있고 차량에 대한 관심을 가진 사람들을 대상으로 모집하는 것이 작업에 대한 이해 및 작업 숙련도를 높이는 데 도움이 될 수 있다.
- 다만, 관련 지식이나 정보에 대하여 너무 전문적인 경우 작업 가이드를 충분히 숙지하지 않고 본인의 지식이나 경험을 가지고 작업을 진행하는 경우도 많기 때문에 이러한 여러 가지 상황을 고려한 작업자 선발 시 우대 조건을 정의해야 한다.

4 선발 방법

- 작업 설계 결과 및 선발 기준을 통하여 필요한 인력을 선발한다. 필요한 인력에 대해서는 전문적인 인력 공급 업체나 아르바이트 모집 사이트 등을 통하여 인력을 확보한다. 이때 선발 기준에 따라 작업자를 선발한다.

4.1 파일럿 프로젝트

- 데이터를 수집하여 정제하거나, 확보된 원천데이터를 작업 기준에 따라서 어노테이션 할 수 있도록 실제 작업 환경을 구성하여 작업과 검사를 진행한다. 이를 통하여 작업과 검수가 가능한 수량을 확인하고 작업의 난이도를 평가하며, 작업 기준과 작업 방법을 설명하는 작업 가이드에 대한 개선 및 보완 작업을 수행한다. 파일럿 프로젝트를 통해서 검수자를 선발하여, 작업자가 가장 용이하게 작업이 가능하도록 작업 기준 및 작업 방법을 보완해야 한다.

4.2 선발 시험

- 작업에 필요한 산업분야 지식이나 작업가이드에 대한 숙지 여부를 평가하기 위하여 선발 시험을 운영한다. 선발 시험의 장점으로는, 불량 작업자 선별 및 결과 확인을 통한 신속한 작업자 선발이 가능하다. 다만 선발 시험의 경우, 예비 작업자들의 부담으로 인한 참여가 저조할 수 있어 주의가 필요하다. 단, 선발 시험 결과를 통해 작업 기준이나 작업 방법 등에 대한 개선 활동을 수행하기에는 한계가 있다.

4.3 면접

- 인력 선발 시 가장 보편적으로 많이 활용되는 방법으로서 모집 인원에 대한 대면 또는 비대면 면접을 통하여 인력을 선발하는 방식이다. 이때 모든 면접관에게 동일한 체크리스트 기반의 면접지를 제공하여 동일한 기준에서 평가될 수 있도록 해야 하며, 작업자와의 면접을 통하여 작업자의 태도나 직업윤리 측면을 확인할 수 있으므로 면접 질문에 관련된 질문이 체크 항목을 포함해야 한다. 면접의 경우 시간과 인력 등 많은 자원이 필요하여 이에 대한 대책 마련이 필요하다.

4.4 서류 심사

- 작업자가 제출한 이력서를 확인하여 선발하는 방법으로 전문 분야의 작업과 관련된 자격증이나 교육 이수 등의 기록을 확인하여 전문 인력을 선발하는 방법이다. 이력서 외에 자격증이나 이수 등의 증명서를 추가로 제출받아 전문 분야 인력임을 확인할 수 있다. 다만, 서류로만 선발하는 경우 가능하면 인공지능 학습용 데이터 구축사업 경험이 있는 작업자 선발을 권장하지만, 이 경우 작업에 대한 성실성이나 전문성을 확인하기 어려운 한계가 있다

제2장 작업자 운영

1 작업 가이드 작성 방법

- 작업가이드는 작업자와 검수자의 작업 방법, 검수 기준 등을 설명한 문서로서 데이터 품질의 기준이 된다. 작업자와 검수자의 작업 품질을 높이기 위해서는 완성도 높은 작업 가이드가 뒷받침되어야 한다. 작업 가이드는 단순히 작업 방법을 안내하는 것을 넘어서 작업자가 작업 과정에서 가질 수 있는 의문과 애로사항들을 스스로 해결할 수 있도록 충분한 정보를 담고 있어야 한다. 작업 가이드가 불분명하거나 정보가 작업자에게 충분히 전달되지 못할 때 작업 품질이 저하될 가능성이 있다.
- 작업 가이드는 작업자에게 혼선을 주지 말아야 한다. 개별 작업자가 동일한 작업의 목표와 의미를 인식하는 과정에서 혼선이 생기면 각자가 판단하는 적절한 작업물의 목표 기준치가 달라진다. 이로 인해 작업 변동성이 커지고, 불필요한 반려와 수정 작업을 반복하게 된다. 따라서 작업 품질을 보증하기 위해 작업 가이드의 작업 방법과 작업 기준 등의 목표와 개념이 모든 작업자에게 일관되고 명확하게 인지될 수 있도록 작업 가이드가 설계되어야 한다.
- 작업가이드는 사전에 미리 정의하지 못한 상황이나 품질검증합의서의 변경으로 인한 경우가 아니면 작업 기준이나 방법이 변경되어서는 안 되며, 작업 기준이나 방법이 변경되면 이전에 작업한 결과물에 대해서 재작업을 통하여 일관성을 유지해야 한다. 다만, 작업 불가나 반려 작업 조건의 경우 작업 중에 주기적으로 업데이트되어야 한다.

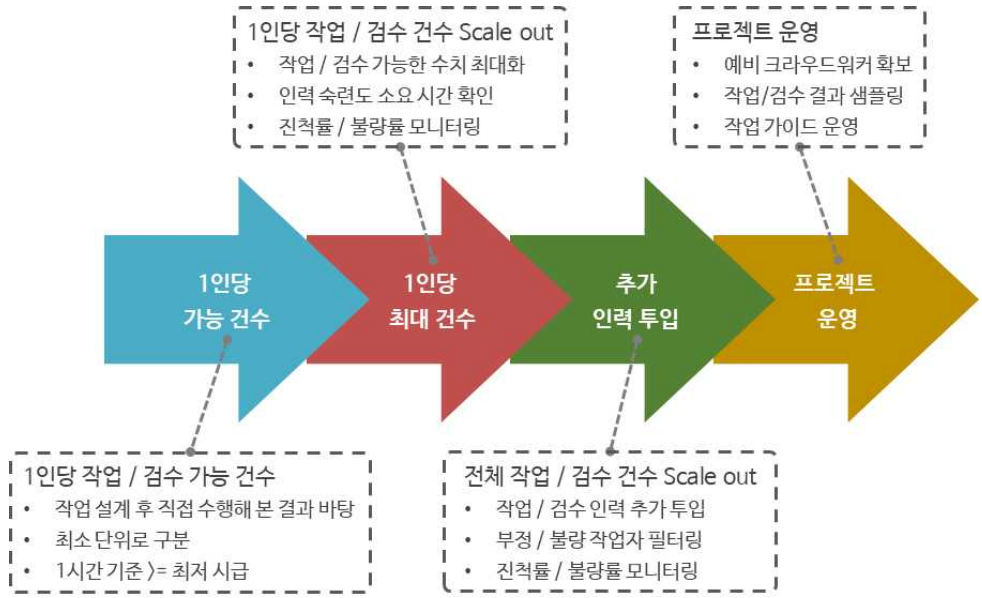
〈표 Ⅰ-1〉 작업 가이드 작성 방법

구분	작업 가이드 작성 방법
정확한 작업 조건과 방법 그리고 수량이 명시되었는지	• 작업자들이 오해하지 않도록 작업 기준에 따라 작업 조건과 방법을 설명하여야 하며, 수량의 경우 정확한 단위에 따른 기준을 명시해야 함 • 이때, 작업자마다 이해와 해석의 차이가 생기지 않도록 작업자가 직관적으로 이해하고 작업할 수 있도록 작성되어야 함

구분	작업 가이드 작성 방법
작업 시 주의사항이 명확하게 제시되고 있는지	• 작업자들이 작업할 때 작업 환경이나 도구 등에서 주의가 필요한 경우 반드시 작업가이드 해당 내용을 명시하고 가능하면 사전 교육을 통하여 주의사항을 명확하게 전달해야 함 • 주의사항의 경우 작업가이드에서 강조하고 교육 등을 통해서 전달하지만 가능하다면 작업 도구에서 시스템적으로 강제할 수 있다면 반영되어야 함
작업 예시 설명과 이미지가 적절하고(제출 가능/불가능) 충분한지	• 작업 예시 설명의 경우 반드시 관련 이미지와 함께 제시되어 작업자과 이해할 수 있도록 해야 함 • 특히, 작업 기준에 해당하는 경우나 예외 상황의 경우는 명확하게 설명과 이미지로 명시해야 함
작업 화면과 조작 방식을 정확하게 설명하고 있는지	• 작업 화면과 조작 방식의 경우 설치가 필요한 경우 설치 파일을 다운로드 받는 것부터 시작해서 설치하고 환경을 설정하고 사용하는 화면 순으로 실제 작업자의 작업 순서에 따라 설명하고, 화면의 해상도와 같이 특정한 값이 있는 경우 환경 설정을 정확하게 할 수 있도록 설명해야 함 • 이때 작업 환경이나 도구가 특정한 버전이나 환경일 때 관련 정보를 명시해야 함
반려 작업 조건이 명확하게 제시하고 있는지	• 작업자가 올바른 작업 기준을 준수하기 위해 작업할 수 있도록 작업 결과물에 대한 검사 결과 반려 기준을 반려 작업에 대한 설명과 이미지를 통해서 명확하게 제시되어야 함 • 반려 작업 조건에 대해서는 작업자가 재작업을 해야 하므로 검수자와 작업자 모두 명확하게 숙지하도록 해야 함
작업 불가 조건이 명확하게 제시되고 있는지	• 작업이 어려운 경우는 작업 대상물에 작업 기준에 해당하는 객체가 존재하지 않거나 작업 방법으로 작업이 불가능한 경우이며, 작업이 매우 어려운 경우에 작업 불가로 판단하는 작업자들도 있기 때문에 작업 불가에 대한 기준이 명확하게 제시되어야 함
직관적으로 이해하기 쉬운 표현을 사용하고 있는지	• 특정 산업분야에서 전문적으로 사용하고 있는 용어·표현이 있다면 해당 용어·표현에 대해서 적절하게 설명이 되어야 하며, 작업가이드에 사용하는 용어나 표현은 중학교 수준에 맞추어 사용하는 것이 적합함
용어는 통일성 있게 사용되고 있는지	• 일관성 있고 명확한 용어가 사용되어야 하며, 작업자가 오해하거나 혼한을 일으킬 수 있는 용어는 배제하며 표준 용어를 사용하여 명확하게 작성되어야 함
사업수행계획서, 품질검증합의서 등의 조건과 일치하는지	• 작업 기준, 방법, 작업 불가 등은 모두 사업수행계획서에서 제안한 수량과 방법을 반영해야 하며, 품질검증합의서의 기준에 따라서 작성되어야 함
작업 수행 시 질의응답 및 연락처 안내가 되어 있는지	• 작업자들이 작업 과정에서 발생하는 문의사항이나 이슈에 대하여 적절한 의사소통과 정보 공유가 가능하도록 관련된 채널이나 서비스 그리고 담당자와 연락처가 명시되어야 함

2 작업 할당

- 작업자의 역량 및 특성에 적합한 작업을 부여해야 한다. 프로젝트 작업의 성격과 작업자의 역량 및 특성 간의 관계를 긴밀하게 파악해 적절하게 작업을 부여하는 것은 전체 프로젝트의 성공의 중요한 요인이다. 특히, 작업자의 성별이나 나이 등 인구학적 속성과 개별 성격에 따라 작업의 품질이 달라지며, 작업자의 개별 특성을 잘 고려하지 못하면 작업 완료된 데이터의 품질이 저하될 수 있다. 따라서, 작업자 1인당 작업 및 검수 가능한 건수를 설계하여 할당하되, 작업 숙련도가 올라오는 경우 최대 가능한 건수를 확인하여 인력을 늘려가면서 작업을 할당해야 한다.



[그림 Ⅰ-2] 작업 할당 절차

2.1 1인당 작업 기준 산정

- 작업 설계 후 직접 수행해 본 결과를 바탕으로 산출하되, 작업 최소 단위로 구분해야 하며, 1시간 기준 작업 및 검수 가능 건수가 최저 시급 이상으로 설계되어야 한다.

2.2 1인당 최대 작업 및 검수 가능 건수 산정

- 작업 / 검수 가능한 최대 수치를 확인하여 인력별 숙련도에 소요되는 시간을 확인하여야 한다. 이를 통하여 1인당 최대 작업 건수 및 검수 가능 평균 건수를 산출하고 이보다

부족한 작업자나 검수자의 경우 진척률과 불량률의 모니터링을 통하여 성실하게 작업을 수행하고 있는지 여부 등을 확인할 수 있다. 또한, 작업자/검수자 별로 숙련도가 평균 이상으로 올라오는데 걸리는 시간을 참고하여 인력에 대한 충원 및 수급 계획을 운영한다.

2.3 전체 작업 및 검수 건수 확대

- 작업 및 검수에 필요한 인력을 처음부터 선발하여 투입하면, 작업이 안정되기 어려운 초기의 경우 확보한 인력들이 작업 없이 대기하는 경우가 많이 발생한다. 따라서 수행기관은 초기에 작업자 및 검수자를 최소한으로 선발하여 운영하면서 필요한 사항들을 확인하고 세팅해야 한다.
- 인력에 대한 추가 투입의 경우 전체 작업 및 검수 건수를 관리가 가능한 수준으로 확대해 가면서 진행하되 부정 작업자나 불량 작업자에 대한 필터링을 하면서 추가 투입된 작업자 및 검수자에 대한 진척률 및 불량률을 모니터링 한다.

2.4 프로젝트 운영

- 일반인을 대상으로 하는 크라우드소싱 기반의 인공지능 학습용 데이터 구축사업의 경우 본인의 사정에 의한 경우도 있고 생각보다 작업이 어렵거나 대가에 대한 불만족 등의 사유로 인하여 일정 비율 이상 작업자의 이탈이 발생하며, 이를 고려한 예비 작업자 확보가 프로젝트 운영 시 중요하다. 또한, 검수 결과에 대한 샘플링 검사를 지속적으로 수행하여 오 검수나 불량 검수로 인한 데이터 품질 저하를 방지해야 한다.
- 그리고, 프로젝트를 운영하면서 작업자나 검수자의 작업에 도움이 될 수 있도록 지속적인 작업 가이드 개선 및 보완 해야 한다. 이때 모호한 기준이나 예외 사항에 대한 업데이트를 통해 좀 더 직관적인 작업 및 검수가 이루어질 수 있도록 한다.

3 작업 모니터링

- 작업자의 작업/검수 정확도와 성실도를 기준으로 작업량, 불량률, 진척률 등을 일·주·월 단위로 모니터링하고 필요한 경우 실시간으로 확인할 수 있도록 관리 체계 수립해야 한다.

3.1 작업량

- 작업량의 경우 근로 형태에 따라 차이가 있을 수 있지만, 작업자별로 작업 건수와 소요 시간 등을 측정하여 작업자의 성실도 등을 확인해야 한다.

〈표 Ⅰ-2〉 작업자 작업량 확인 지표

지표	내역
작업 건수	• 개인별, 전체, 누적, 평균 작업 건수 등을 측정하고 관리해야 함
작업 소요 시간	• 작업 건수별로 실제 작업에 투입되는 시간으로서 개인별 작업 소요 시간, 평균 작업 소요 시간 등을 측정하고 모니터링 해야 함
작업자수	• 근로 형태에 따라서 차이가 있지만 아르바이트의 경우 참여한다고 하면서 실제로 작업을 하지 않는 경우도 있어서 전체 계약 인원과 실제 작업 인원이 다를 수 있으므로 구분하여 관리함
검수 대기 건수	• 작업자의 작업 결과에 대한 검수 처리가 되지 않고 남아 있는 작업 결과물 건수를 측정하여 검수 완료 건수와 함께 관리하여 검수자의 성실도 등을 확인하고 관리함
검수 완료 건수	• 검수자가 완료 처리한 작업 건수로서 실제 완료한 작업 건수로 측정하고 관리함
검수 소요 시간	• 검수자의 경우 작업 결과물이 작업 기준에 맞춰서 제대로 작업이 이루어졌는지를 검사하는 역할로서 검수 소요 시간이 평균 보다 낮은 경우 검수를 제대로 하지 않은 상황으로 추정할 수 있으며, 시간이 너무 많이 소요되는 경우 검수자로서 역량이나 숙련도에 문제가 있을 수 있는 상황으로 추정하여 관리한다. 이를 위하여 검수자별 검수 소요 시간, 전체 검수 평균 시간, 건당 검수 시간 등을 측정하여 모니터링 함
검수자수	• 계약한 인원과 검수 완료 건수 등을 참고하여 실제 검수 작업을 수행하는 인원으로 구분하여 모니터링 함

3.2 진척률

- 작업량과 관련된 지표는 주기적으로 모니터링해야 한다. 가능하면 실시간으로 모니터링할 수 있는 경우가 가장 바람직하지만 실시간 모니터링이 어렵다면 최소 일, 주 단위의 모니터링을 통하여 작업량과 지표를 확인해야 한다.

〈표 Ⅰ-3〉 작업자 작업 진척률 확인 지표

지표	내역
1일 평균 작업/검수 건수	• 전체 작업에 대해서 평균 건수가 사업 초기에는 증가하다가 작업자들의 숙련도가 높아지면 일정한 건수로 안정화되며 이러한 평균 작업/검수 건수의 변화를 통하여 매일 평균 작업/검수 건수를 확인하고 이를 통하여 전체 작업자들의 성과와 사업 수행 간의 이슈를 확인할 수 있음

지표	내역
작업/검수 소요시간	• 작업/검수에 소요되는 시간을 확인하여 전체 작업자들의 작업/검수에 대한 집중 여부나 난이도 등을 확인할 수 있음
1인 작업/검수 건수	• 작업자 1인별로 작업/검수 건수를 모니터링 하여 평균보다 미흡한 경우 재교육이나 작업 할당 조정 등을 통하여 관리하고 일정 기간 이상 평균 건수에서 미흡한 경우 불성실 작업자로서 작업에서 배재하는 등으로 관리하여야 한다. 특히, 검수자의 경우 검수 건수가 평균보다 과대하게 초과한 경우 검수를 불성실하게 하였을 수 있기 때문에 해당 검수자의 검수 결과를 전수 검사하여 검수 작업에서 배제하는 등으로 관리하여야 함
예상 완료율	• 작업 건수와 작업 시간 등에 대해서 안정화되면, 기간, 투입 인력 등을 고려한 예상완료율을 산출할 수 있으면 이를 통하여 작업 관리를 수행해야 함 (예상완료율 = 남은 기간 x 투입 인원 x 1일 평균 작업 건수/검수 건수)

3.3 불량률

- 불량률은 총 작업량에서 관찰된 총 결함 작업량의 비율을 의미한다.
- 작업자와 검수자에 대해서 작업 기준 통과 여부를 불량의 기준으로 삼아 모니터링하고 이를 기반으로 데이터 품질을 높여야 한다.

〈표 Ⅰ-4〉 작업자 작업 불량률 확인 지표

지표	내역
전체 반려 건수/반려율	• 전체 진행되고 있는 작업의 검수 결과 반려 건수와 반려율을 측정하여 관리하여야 한다. 이를 통하여 전반적인 작업의 숙련도, 난이도 등을 추정할 수 있으며 반려 건수와 반려율의 증감에 재교육, 작업가이드 반영, 인력 필터링 등을 수행해야 함
반려가 많은 작업자	• 작업자들의 반려 건수와 반려율에 대해서 평균/최대/최소를 구분하여 모니터링 하여 반려 건수와 반려율이 높은 작업자의 경우 작업 숙련도가 높지 않거나 작업에 대한 이해가 부족하거나 작업을 불성실하게 하는 경우 등이므로 재교육이나 작업 할당 조정 등을 통하여 관리하고 개선이 되지 않는 경우 작업에서 배제할 수 있음
반려가 적은 검수자	• 검수자의 반려 건수와 반려율에 대해서 평균/최대/최소를 구분하여 모니터링 하여 반려 건수와 반려율이 낮은 검수자의 경우 검수 기준에 대한 이해가 부족하거나 검수를 불성실하게 하는 경우 등이므로 재교육이나 작업 할당 조정 등을 통하여 관리하고 개선이 되지 않는 경우 검수 작업에서 배제할 수 있음

3.4 샘플링

- 샘플링은 모집단에서 일정한 수만큼 추출하는 작업을 말한다. 개별 관측치의 선택과 관련된 통계적 절차로서, 모집단에 대한 통계적 추론을 하는 데 도움이 된다. 샘플링을 통해서 작업자와 검수자에 대한 사후관리가 가능하다. 샘플링 결과 불량이면 해당 작업자/검수자

의 결과물을 전수 검사하여 작업자의 작업 방법과 작업 기준에 대한 이해를 확인하여야 한다.

- 상호평가와 사후관리를 하는 과정은 작업자와 검수자의 학습효과를 강화한다. 작업자 간, 혹은 검수자 간 상호평가를 통해 공유함으로써 데이터의 품질을 향상시킬 수 있다.

제3장 작업자 관리

1 관리

- 작업자에게 적절한 동기부여와 원활한 소통을 진행하되 Micro Managing을 통해 수시로 부정/불량 작업자를 관리해야 한다. 인공지능 학습용 데이터 구축사업에서 품질 관리는 작업자 관리에서 시작된다고 할 수 있다.

1.1 불성실 작업자 & 중간 이탈자 관리

- 인공지능 학습용 데이터 구축사업에서 발생할 수 있는 불성실 작업자와 중간 이탈자를 방지할 수 있는 대안을 마련해야 한다. 불성실 작업자와 이탈자의 비율이 높아질수록 작업 완료에 필요한 소요 시간이 증가하고, 데이터의 품질이 저하되기 때문에 이들에 대한 효과적인 관리가 필요하다.
- 작업 과정에서 함정 질문을 삽입해 작업자의 참여 의지와 부정 작업 여부를 판별할 수 있다. 여기서 삽입되는 질문은 작업과 연관성은 떨어지지만, 작업에 주의를 기울이지 않거나 매크로를 설정하는 등의 부정 작업 여부를 식별하는 데에는 유용하게 활용될 수 있다. 따라서 중간 질문은 누구나 정답을 제시할 수 있는 상식적인 퀴즈 형식(예시 : 대한민국의 수도는 어디인가?)으로 삽입하여 활용한다.
- 이전 장에서 설명한 작업건수, 반려 건수, 반려율 등으로 구성된 작업량, 불량률 등과 같은 작업과 관련된 지표를 만들어 작업자를 모니터링하고 관리해야 한다. 각 요소의 정의와 측정 방식을 규정하고, 요소를 계산하는 데 필요한 지표를 정한다. 각 지표에 따른 불량 작업자 기준을 세우고, 이를 프로젝트 수행 시 모니터링에 사용하여 불량 작업자를 식별한다. 1차적으로 불량 작업자로 판별된 작업자들은 바로 퇴출하지 않고, 지속적인 추적 관찰을 통해 걸러내도록 한다.

- 중간 이탈자를 방지하기 위하여 작업자를 독려하고 작업 공백에 유연하게 대처할 수 있는 작업 환경을 구성해야 한다. 일반적으로 더 많은 보상을 주는 것이 작업자의 근로의욕을 높이는 데 도움이 되지만 개별 작업자의 효용 곡선을 일일이 파악하는 것이 불가능하기 때문에 성과 대비 장기적인 비용 부담을 초래하기 쉽다. 따라서 추가적인 금전적 보상을 제공하는 것에 치중하기보다는 작업자의 내적 동기를 높이는 방안들을 강구하고 이탈에 대응할 수 있는 프로세스를 설계할 필요가 있다.
- 작업자의 내적 동기를 강화하기 위해서는 작업에 대한 책임감과 작업팀에 대한 소속감 및 연대감을 가질 수 있는 장치를 마련해야 한다. 작업자 환경에서 형성되는 노동환경의 특성은 대체로 일회성이 강하며 작업자들 또한 일용직으로 인식하는 경우가 많은 게 사실이다. 그렇다 보니 단위 작업에 대한 책임 의식과 프로젝트 참여자들 간의 연대 의식이 약화 될 수밖에 없다.
- 프로젝트 관리자는 이에 대한 대응책으로 사전에 개별 작업자에 대한 최소 작업 요구치를 설정하고 해당 작업의 명확한 목표와 전체 프로젝트에서 가지는 의미를 지속적으로 각인시킬 수 있어야 한다. 여기서 개별 작업의 목표를 설정하는 경우 내용이 구체적이고 난이도가 적당하며, 작업자가 납득할 수 있는 내용인지를 고려해야 한다.
- 명확한 목표의식을 형성하는 것은 작업자 스스로가 작업에 대한 책임감과 개별 기여의 중요성을 느낄 수 있게 한다. 실제로 작업자가 스스로 의미 있는 작업을 수행한다고 인식할 때, 중간 이탈자가 감소하는 경향을 보인다.
- 이와 더불어 사전에 합의된 최소 기준을 충족하지 못하거나 동료 작업자에게 피해가 가는 무책임한 태도를 보일 경우 부여될 수 있는 페널티 (기존 보상회수, 타 프로젝트 참여 제한 등)와 함께 법적 제재의 가능성을 공시해야 한다. 이를 통해 무분별한 이탈을 방지함과 동시에 이탈로 인한 전체 프로젝트의 피해를 예방할 수 있다.
- 프로세스 측면에서는 예기치 못한 작업 공백을 메꿀 수 있는 온디맨드(On-demand) 형식의 추가 매칭 프로세스를 구현해야 한다. 특히 프로젝트를 진행하는 과정에서 충분한 작업 참여 의지와 숙련도를 갖췄음에도 불구하고 선발 기회를 얻지 못하거나 후 순위로 밀려난 작업자들이 생길 수 있는데, 관리자는 이들을 효율적으로 활용할 수 있어야 한다. 이를 위해 사전 선별 및 선발 과정에서 채택되지 못한 예비 작업자들의 명단을 확보하고 추가 투입 가능 여부를 지속적으로 파악하는 작업이 필요하다.

1.2 작업자 소통

- 크라우드 소싱 환경에서는 프로젝트에 관한 작업자와 프로젝트 관리자 간의 정보비대칭을 완화해야 한다. 프로젝트에 관한 정보와 권한이 관리자 직급에 집중되면 작업자들의 주인의식이 희석되고 불신이 커질 수 있다. 반대로 작업자가 전체 프로젝트의 현황 (진행률, 반려율, 작업자별 기여도 등)을 파악할 수 있고 동료 작업자 및 프로젝트 관리자와 원활하게 소통하며 정보를 공유하는 환경에서 작업 동기부여가 향상될 수 있다. 특히 다수의 작업자가 참여하는 협업 프로젝트의 경우 구성원들 간의 소통이 원활하고 공유할 수 있는 정보와 권한의 범위가 확장될 때 역량이 극대화될 수 있다.

2 교육

- 작업자와 검수자의 작업 역량을 개선하고 작업의 신뢰성을 확보하며 데이터 품질을 높이기 위해서는 작업자와 검수자를 대상으로 적절한 교육을 진행해야 한다. 작업자와 검수자를 대상으로 한 교육은 작업 수행 이전뿐만 아니라 작업 과정 전반에 걸쳐 연속적으로 이루어져야 한다. 예비 작업자를 대상으로 한 사전 교육은 인공지능 학습용 데이터에 대한 기본적인 이해와 더불어 데이터 라벨링의 목적과 구체적인 작업 방법 및 가이드를 명확하게 인지할 수 있도록 설계되어야 한다. 아무리 특정 작업 산업분야에 대한 지식이 많은 작업자라도 작업 목표가 불명확하거나 작업 방법에 익숙하지 않으면 데이터의 품질이 저하된다.
- 작업자 교육 방식에는 작업에 대한 참값과 훈련용 데이터 등으로 준비된 샘플 데이터에 대한 작업을 통하여 작업자를 훈련하는 방식과 다른 작업자의 작업 결과물을 검토해나가며 학습하는 방식이 있다. 전자의 경우 학습효과 측면에서 작업자의 정확성을 높이는 데 가장 유용한 방식이지만 사전에 훈련을 위한 샘플 데이터를 구축하는 등의 비용이 발생할 수 있으며, 후자의 경우 별도의 샘플 데이터 구축이 필요 없기 때문에 소요 비용이 감소하고 효과적으로 작업 능률이 향상된다는 장점이 있다.
- 검수자는 작업자와 다른 역량 및 경험을 요구하기 때문에 독립적이고 전문화된 교육 과정을 마련해야 한다. 특히, 검수자가 작업 결과물에 대한 적합성을 판별하고 객관적이고 중립적인 기준을 설정할 수 있도록 교육하여야 한다. 또한, 작업자와 검수자의 교육 결과의 검증을 위해서 숙련도 및 적합도를 측정하는 평가제도 및 자격시험을 도입할 수도

있다. 평가 제도는 요구기준에 미달 되거나 부정 작업자 및 검수자를 선별하여 제거함으로써 작업자 및 검수자의 품질에 긍정적인 영향을 미친다. 그리고 작업자의 개별 수행 능력을 공식적으로 입증할 수 있는 자격을 부여하는 성격을 가진 자격시험을 도입하면 작업자의 능률과 의욕을 높일 수 있다.

3 프로젝트 진척률 제고

- 작업 평가 기준과 보상 책정 기준을 명확하게 설정하고 이를 작업자에게 사전에 공지해야 한다. 작업자가 쉽게 이해하고 납득할 수 있는 기준이 마련될수록 그에 맞는 작업을 수행할 가능성이 높아지기 때문이다. 금전적 보상의 경우 보상의 절대적인 양보다 어떠한 기준에 따라 보상이 책정되고 지급되느냐가 결과 품질에 더 큰 영향을 미친다. 따라서 프로젝트 관리자는 기준을 설정하고 실행함에 있어 분배과정과 보상이 공정하게 이루어질 수 있도록 투명하게 운영하여야 한다.

3.1 보수, 인센티브

- 작업자는 보수, 인센티브에 크게 반응한다. 대체로 보수와 인센티브가 높을수록 더 많은 일을 처리하는 것으로 나타났다. 또한 작업자는 같은 시간이 주어졌을 때 더 많은 보수를 제공하는 프로젝트에 참여하고 싶어 한다. 프로젝트 관리자는 이러한 측면들을 고려해 적절한 보수와 인센티브 시스템을 마련해야 한다. 관리자는 전체 프로젝트의 진행성과가 미진하다면 인센티브를 올리는 방법을 택할 수 있고, 작업 난이도가 너무 높을 경우 보수를 높게 측정해 작업자를 유인할 수도 있다. 다음은 수행기관이 크라우드소싱 기반의 작업을 진행하는 동안 얻은 사례를 정리해놓은 것이다.

3.2 보수와 작업완료 속도의 관계

- 우선 작업 당 제공되는 보수가 높을수록 작업자가 작업을 완료하는 속도가 빨라질 수 있었다. 같은 난이도의 작업을 진행하는 프로젝트라고 가정했을 때, 작업자의 보수를 높인다면 작업을 더 빠르게 완료할 수 있었다. 다만 수행기관은 정해진 예산 안에서 프로젝트를 완료해야하기 때문에 시간 절약만의 이유로 작업자에게 제공하는 보수를 높이기는 쉽지 않았다. 이미지 가공 프로젝트를 진행할 때, 한 장당 보수를 100원으로 책정했을 때와

비교하여 200원으로 책정했을 때 수행기관이 부담해야하는 금액의 차이는 배로 늘어났다. 수행기관은 보수를 추가 지급하는 비용과 프로젝트 시간 절약으로 얻는 효용을 비교해 적절한 전략을 선택해야 한다.

- 단순히 작업자에게 제공하는 보수를 처음부터 높게 책정하여서 프로젝트를 빠르게 완료시킨다는 전략보다, 프로젝트의 진척도가 미진하다고 판단될 때 작업자들에게 완성 속도와 결과 품질의 경과에 따른 인센티브를 제공하는 것이 더 효율적일 수도 있다.

3.3 보수와 작업완료 양의 관계

- 작업자에게 제공되는 보수가 높을수록 완료되는 작업의 양은 유의미하게 늘어날 수 있었다. 이는 작업 난이도와는 별개로 보수를 높일수록 작업자는 일을 빠르게 완료하고, 작업자가 전체 작업 기간 동안 완료한 일의 양은 크게 증가하였다.

3.4 보수와 작업 품질의 관계

- 한편 데이터 라벨링 작업의 난도가 증가하면 작업의 품질은 유의미하게 감소했다. 또한 작업자에게 지급되는 보수가 높아진다고 난도가 높은 작업의 품질이 좋아지지 않았다.
- 작업자들은 대체로 데이터 라벨링 작업을 하면서 어려운 난도의 작업은 기피하는 경향이 있다. 또한 개인의 데스크톱, 노트북 PC, 또는 모바일 핸드폰을 이용해서 작업을 진행하기 때문에 작업자는 작업에 싫증을 느끼면 언제든지 작업을 그만둘 수 있다. 프로젝트 관리자는 프로젝트에서 수행할 작업을 적절한 난이도로 분류해서 작업자에게 할당해야 한다. 만약 작업의 난도가 너무 높을 시에는 작업자들의 작업 참여를 유도하기 위해 높은 보수를 책정하고, 내재적 동기를 이끌어내야 한다.

3.5 보수와 작업자 내재적 동기

- 프로젝트 관리자는 보수와 인센티브를 통해서 작업자의 내재적 동기를 이끌어내야 한다. 작업자들은 자신이 하고 있는 일이 의미 있는 일이라고 생각할 때 더 높은 품질의 작업 결과물을 보여줬고 완료한 작업의 양도 증가했다. 이에 관리자는 작업자에게 현재 하고 있는 작업의 가치를 작업자들에게 지속적으로 상기시켜 줄 필요가 있다. 많은 작업자들은 크라우드소싱 작업을 단순 반복적인 업무가 데이터 라벨링 작업의 전부라고 생각하는 것이다. 하지만 앞서 언급했듯이, 외국에서는 크라우드소싱 작업을 통해 탄탄한 줄거리를

갖춘 소설도 창작했다. 프로젝트 관리자가 적절한 업무 프레임워크를 갖춘 작업을 설계한다면 창의적인 일도 수행할 수 있다. 이와 같은 해외의 선진 크라우드소싱 사례를 소개하며 작업자들의 내재적 동기를 이끌어낼 수 있다. 또한 작업자는 자신이 작업 당 받는 보수가 자신이 하는 일의 가치라고 생각한다. 작업에 대한 보수가 높을수록 자신이 하는 일은 실제로 가치가 있는 일이라고 생각하는 것이다. 작업자의 작업에 대한 내재적 동기를 최대한으로 끌어올리고 싶다면 작업의 보수 혹은 인센티브를 높이는 것도 좋은 방법이다.

3.6 보수 지급 기준의 일관성 유지

- 작업자들에게 작업 보수 지급 기준을 투명하게 제시하고 일관되게 유지하는 것 또한 중요하다. 보수 지급 기준이 정해져있지 않거나 처음에 제시되었던 기준과 다르게 보수가 지급되었을 때 작업자들은 작업을 불성실하게 수행하였다. 예를 들면, 앞으로 데이터 라벨링 작업을 적극적으로 참여할 MZ세대(1980년대~2000년대 중반 태생)는 공정성과 투명성을 중요하게 생각한다. MZ세대 작업자들의 특성을 고려한다면, 보수 지급 기준을 투명하게 제시하고 일관되게 유지하는 것은 당연히 해야 하는 일이다.

3.7 세부 작업에 대한 보수 결정

- 난이도와 복잡성이 높은 작업의 경우 세부 작업으로 작업을 세분화하고 작업흐름을 체계화한다면 작업자들의 참여의욕을 높일 수 있었다. 이를 고려하여 프로젝트 관리자는 프로젝트의 작업을 쪼개고 어떤 작업 단위별로 보수를 지급할지 결정해야 한다. 세부 작업 단위의 크기(Granularity)는 작업 품질에 유의미한 영향을 미칠 수 있으며, 각 단위를 연결하는 프로세스와 작업 단계를 적절하게 설계했을 때 전체 데이터 품질을 보장할 수 있다.

3.8 단가 책정에서의 고려사항

- 프로젝트 관리자는 작업 단가와 보수를 설정할 때 작업자의 의사를 파악하고 이를 반영해야 한다. 관리자는 현실적으로 자신의 관점에서 용역의 가치를 결정하게 되므로 작업자가 단가와 보상 규정에 대해 불만을 갖게 된다면, 참여 의욕이 저하되거나 중도에 이탈하여 전체 품질에 악영향을 미칠 수 있다. 이에 대한 대안으로 프로젝트 관리자가 단가의 상한선(Threshold Price)을 설정하고 작업자 주도로 경매(Bid)를 진행함으로써 단가를 책정하는 방식을 활용할 수도 있다. 이를 통해 작업자의 이윤 추구 적 행동을 통해 기존에 파악하

기 어려웠던 작업자의 개별 효용을 반영하고 단가에 대한 불만을 줄일 수 있다. 그러나 무분별한 경매를 방지하기 위하여 관련 규정을 구체화할 필요가 있으며, 단가 책정이 어렵거나 작업자의 불만이 많은 경우에만 부분적으로 활용하는 것이 효과적이다.

Ⅱ
공통참조항목

제1장 개요

1 작성 배경

- 인공지능 학습용 데이터의 구축 수행 및 참여기관에 따라 라벨링 데이터의 유형(xml, json 등) 및 구조 등이 상이하여 기 구축된 인공지능 학습용 데이터를 다른 목적으로 활용하기 어렵고, 구축된 데이터 품질 관리 문제가 발생
- 인공지능 학습용 데이터 라벨링 공통 기준을 통해 원시데이터 유형과 목적별 라벨링 구조 및 포맷 등의 기준이 필요함

2 작성 목적

- 인공지능 학습용 데이터 구축사업의 수행 및 참여기업에서 인공지능 학습데이터의 수집 · 정제 · 라벨링 절차에 따라 구축 목적과 원천데이터 유형에 맞는 구축사업에서 참조할 수 있는 공통사항들을 위한 기준 마련
- 인공지능 학습용 데이터 구축사업에 참여하는 신규 수행 및 참여기관의 라벨링 구축 방식을 안내하는 것을 목적으로 인공지능 학습용 구축사업 시 사업계획서에 사전 반영 및 배포하여 사용하고자 함
- 인공지능 학습용 구축사업의 인공지능 기술 분야에 대한 라벨링 관련 공통참조기준 제공을 통해 고품질의 인공지능 학습용 데이터 셋을 확보하고자 함

3 작성 범위

- ’20년 인공지능 학습용 데이터 구축사업 8개 영역 48개 분야별 150종 인공지능 학습용 데이터 획득 공통 참조기준 가이드라인과 기존 TTA 인공지능 학습용 데이터 구축공정 가이드라인을 분석 검토하여 최소한의 기준으로 공통 참조 가능한 항목을 수립하여 도출
- 해외 인공지능 주요 사례에서 확인 후 라벨링 방식 별 데이터 구조 참조
- 국내 학습데이터 관련 인공지능 기업 사례 조사 등 다양한 기업들의 선행 사례들을 확인 후 가능 방법 적용
- 인공지능 학습용 데이터 구축사업 중 텍스트와 광학문자인식 이미지, 자율주행, 영상(동적/정적)이미지 총 4종 영역을 분석하여 최소한의 기준으로 공통 활용 가능한 기준을 수립하여 도출

3.1 공통참조항목 도출 방법

- ’20년 인공지능 학습용 데이터 구축사업 1차 구축 영역 중 관련 결과물 분석을 통한 라벨링 구조 도출
- 한국어 분석에 가장 많이 사용되는 한국전자통신연구원(ETRI)의 KorBERT 언어모델을 중심으로 SKT의 KoBERT 언어모델과 HanBERT 언어모델을 분석하여 텍스트 유형 공통 참조기준 도출
- Google Vision API와 Tesseract OCR을 참조하여 이미지 유형 중에서 이미지 내 텍스트 추출(OCR)에 대한 공통참조기준 도출
- 자율주행 관련 ’20년 인공지능 학습용 데이터 1차 구축사업 영역 결과물 분석 및 관련 기업의 라벨링 자료 획득 및 분석을 통한 자율주행 유형 공통참조기준 도출
- 영상(동적/정적)이미지의 경우 ’20년 인공지능 학습용 구축사업 1차 20개와 2차 150개를 확인 후 종별 101개 영역을 분석 후 획득 방식에 따른 분류 작업 진행
- 영상(동적/정적) 이미지 가운데 획득 방법에 대한 장비별 분류를 통하여 렌즈와 해상도가 동일할 수 있는 장비와 묶어서 분석 검토

- ’20년 추경 150종 영상(동적/정적) 이미지 내용 분석(디지털 일안 반사식 카메라(Digital Single Lens Reflex Camera 이하 DSLR) 또는 디지털 카메라, 스마트폰, 드론/위성, 폐쇄회로 TV(Closed-circuit television 이하 CCTV), 특수촬영 장비 사용)
  - 광학 문자 인식(Optical Character Recognition 이하 OCR), 헬스케어(X-Ray, 자기공명영상(Magnetic Resonance Imaging 이하 MRI), 컴퓨터단층촬영(Computed Tomography 이하 CT), 특수렌즈(열화상, 초분광) 사용 카메라 등은 검토 및 확인하고 공통 참조 항목은 추가 사업을 통해 진행 예정
- 각각의 영상(동적/정적) 이미지 획득 방법과 공통 사용 항목 도출 방식
  - 개개별 과제 내 영상 속성 정보 확인
  - 공통 참조분모 활용 기준 도출
  - 수행기관이 제출한 구축공정 활용 가이드라인에서 수집 대상과 방법 검토 확인
  - 국내외 영상(동적/정적) 이미지 획득 사례 참조

4 용어 정의

- 데이터 수집 (또는 획득) (Data Acquisition)
  - 인공지능의 기계학습에 필요한 데이터를 현실 세계에서 직접 수집 또는 생성하거나, 이미 보유하고 있는 조직이나 시스템 등으로부터 법률적 제약이 없도록 ‘원시데이터’를 확보하는 활동
- 데이터 정제 (Data Refinement)
  - 획득한 원시데이터를 기계학습에 필요한 형식으로 맞추거나 불필요한 중복을 제거하며, 개인정보를 비식별화하여 처리하는 등 일련의 전처리 과정을 통해 ‘원천데이터’를 확보하는 활동
- 데이터 라벨링 (또는 가공) (Data Labeling)
  - 인공지능이 기계학습에 활용할 수 있도록 기능이나 목적에 부합하는 정보를 원천데이터에 부착하는 활동

- 라벨링 데이터 (Labeled Data)
  - 원천데이터에 부여한 '참값', 파일형식이나 해상도 등의 속성, 그리고 설명이나 주석 등이 포함된 '어노테이션'의 집합
- 원시데이터 (Raw Data)
  - 기계학습을 목적으로 획득 단계에서 수집 또는 생성한 음성, 이미지, 영상, 텍스트 등의 데이터
- 원천데이터 (Source Data, Unlabeled Data)
  - 원시데이터를 라벨링 공정에 투입하기 위해 필요한 전처리 등 정제 작업을 수행한 데이터로 라벨링데이터가 부여되지 않은 상태의 데이터
- 인공지능 학습용 데이터 구축
  - 임무정의, 데이터 획득, 데이터 정제, 데이터 라벨링 등 인공지능 학습용 데이터를 구축하는 일련의 활동
- 참값 (Ground Truth)
  - 인공지능의 기계학습 목적에 따라 원시데이터에 라벨링된 정확한 값이나 사실의 의미적 표현
- 어노테이션 (Annotation)
  - 데이터 라벨링 시 원시데이터에 주석을 표시하는 작업을 의미하며, 추가 부착되는 설명정보 데이터는 기능 목적에 따라 다양한 형태로 표현될 수 있으며 이러한 설명정보 표현방식을 지칭

    ※ 용어사용 예 : 사물 바운딩박스 어노테이션, 클래스 라벨링 어노테이션 등
- 광학문자인식 (OCR, Optical Character Recognition)
  - 사람이 쓰거나 기계로 인쇄한 문자의 영상을 기계가 읽을 수 있는 문자로 변환하는 것

제2장 인공지능 학습용 데이터셋 구축 공통참조항목

1 원시데이터 유형별 라벨링 기능 및 어노테이션 방식

〈표 Ⅱ-1〉 원시데이터 유형별 라벨링 기능 및 어노테이션 방식

No	데이터 유형	라벨링 기능	어노테이션 방식
1	텍스트	• 텍스트 분류(Text Classification)	• 클래스 라벨 (단일, 다중)
		• 개체명 인식(Named Entity Recognition)	• 단어(구문) 라벨
		• 관계-의존성 정의(Relation-Dependencies)	• 단어(구문) 라벨링 및 두 단어 사이의 관계
2	이미지	• 이미지 분류(Image Classification)	• 클래스 라벨 (단일, 다중)
		• 객체 인식(Object Recognition)	• 바운딩 박스(사각형) • 폴리곤(다각형)
		• 영역 구분(Segmentation)	• 픽셀(점)
3	비디오	• 동영상 분류(Video Classification)	• 클래스 라벨 (단일, 다중)
		• 객체 인식(Object Recognition)	• 바운딩 박스(사각형) • 키 포인트(정점) • 폴리곤(다각형) • 폴리라인(선)
		• 객체 추적(Object Tracking)	
4	오디오	• 오디오 분류(Audio Classification)	• 클래스 라벨
		• 오디오 세그멘테이션(Audio Segmentation)	
		• 음성인식(음성 → 텍스트 변환) (Speech to Text)	• 텍스트 전사
5	기타	• 시계열 세그멘테이션 (Time-Series Segmentation) • HTML 문서 분류(HTML Classification)	• 클래스 라벨

2 텍스트 데이터

2.1 라벨링 공통 항목

- 텍스트 유형의 인공지능 학습용 데이터는 '문서요약', '질의응답', '기계번역', '대화' 등 다양한 목적으로 구축되며, 가장 기본이 되는 언어 분석을 위한 라벨링 구조를 공통기준으로 도출

〈표 Ⅱ-2〉 라벨링 메타정보 공통참조항목

No	속성명	항목 설명	Type	필수여부	작성예시
1	Dataset.identifier	데이터셋 식별자	string	필수	TEXT_QnA_LAW_01 (데이터유형_목적_분야_순번)
2	Dataset.name	데이터셋 이름	string	필수	법률 관련 인공지능 질의응답 학습용 데이터 셋
3	Dataset.src_path	데이터셋 폴더 위치	string	필수	/dataSet/text/
4	Dataset.label_path	데이터셋 레이블 폴더 위치	string	필수	/dataSet/text/
5	Dataset.category	데이터셋 카테고리	number	필수	0: 텍스트 분류, 1: 문서요약, 2:질의응답, 3: 기계번역 등
6	Dataset.type	데이터셋6 타입	number	필수	0: 텍스트, 1: 이미지, 2:영상, 3: 음성 등

- '문서요약' 및 '문서분류' 등의 목적으로 원시데이터의 출처 정보가 중요한 경우 선택적으로 작성

〈표 Ⅱ-3〉 라벨링 메타정보 선택항목 - 1

No	속성명	항목 설명	Type	필수여부	작성예시
1	info.filename	원시데이터 파일명	string	선택	NEWS_000001 (매체유형_순번)
2	info.title	원시데이터 제목	string	선택	이스라엘 75세 남성 화이자 백신 접종 후 사망… "백신 연관성 없는 듯"
3	info.mediatype	매체유형	string	선택	뉴스, 블로그, SNS 등
4	info.medianame	매체명	string	선택	중앙일보

No	속성명	항목 설명	Type	필수여부	작성예시
5	info.category	원시데이터 카테고리	string	선택	정치, 경제, 연예, 스포츠 등
6	info.size	원시데이터 크기 (글자수)	string	선택	270
7	info.date	발행일자	string	선택	2020.12.29 12:40:23 (yyyy.MM.dd HH:mm:ss)

- 저작권 정보가 존재할 때, 선택적으로 작성

〈표 Ⅱ-4〉 라벨링 메타정보 선택항목 - 2

No	속성명	항목 설명	Type	필수여부	작성예시
1	licenses.id	라이센스 고유 번호	string	선택	http://www.apache.org/ licenses/LICENSE-1.0
2	licenses.name	라이센스 이름	string	선택	Apache License 1.0
3	licenses.url	문서 식별자	string	선택	NEWS_000001

2.2 어노테이션 공통 항목

- 자연어의 의미 분석을 위해 문장 단위로 형태소 분석을 위한 어노테이션 항목

〈표 Ⅱ-5〉 자연어 처리 어노테이션 항목 - 1

No	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].morp .id	형태소 식별자 (출현 순서)	string	선택	NNG_00001 (형태소태그_순번)
2	annotations[].morp .lemma	형태소	string	선택	일반명사
3	annotations[].morp .type	형태소 태그	string	선택	NNG, NNP, NNB 등
4	annotations[].morp .position	문장 내 위치	number	선택	231 (형태소 위치)
5	annotations[].morp .weight	형태소 분석 결과 신뢰도	number	선택	0.92 (0 ~ 1)

- 자연어의 의미 분석을 위해 단어 혹은 구문 단위 분석을 위한 어노테이션 항목

〈표 Ⅱ-6〉 자연어 처리 어노테이션 항목 - 2

No	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].text_id	원문 텍스트 식별자	string	선택	TXT_0001 (분류_순번)
2	annotations[].word[] .string	단어	string	선택	네이버
3	annotations[].word[] .label	단어 레이블	string	선택	기업
4	annotations[].word[] .start	원문 내 단어 시작 지점	number	선택	100
5	annotations[].word[] .end	원문 내 단어 종료 지점	number	선택	102

- 자연어의 의미 분석을 위해 문장 단위로 어휘의미 분석을 위한 어노테이션 항목

〈표 Ⅱ-7〉 자연어 처리 어노테이션 항목 - 3

No	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].WSD .id	어휘의미 식별자 (출현 순서)	string	선택	WSD_0001 (분류_순번)
2	annotations[].WSD.t ext	어휘 텍스트	string	선택	배
3	annotations[].WSD. weight	어휘의미 분석 결과 신뢰도	number	선택	0.92 (0 ~ 1)
4	annotations[].WSD. position	문장 내 위치	number	선택	2310
5	annotations[].WSD. begin	어휘의 첫 형태소 식별자	string	선택	LC_0012
6	annotations[].WSD. end	어휘의 끝 형태소 식별자	string	선택	LC_0017

● 텍스트 데이터 라벨링 기능 및 어노테이션 방식

〈표 Ⅱ-8〉 텍스트 데이터 라벨링 기능 및 어노테이션 방식

No	라벨링 기능	어노테이션 방식
1	• 텍스트 분류(Text Classification)	• 클래스 라벨(단일, 다중)
2	• 개체명 인식(Named Entity Recognition)	• 단어(구문) 라벨
3	• 관계-의존성 정의(Relation-Dependencies)	• 단어(구문) 라벨링 및 두 단어 사이의 관계

● 클래스 어노테이션 예시

〈표 Ⅱ-9〉 클래스 어노테이션 예시

No	속성명	항목 설명	Type	필수여부	작성예시
16	annotations[].id	어노테이션 식별자	string	선택	CL_0001 (분류_순번)
2	annotations[].class	클래스 분류 (클래스 정의 필요)	number	선택	0: 정치, 1: 사회, 2: 연예, 등

● 텍스트 내 개체명 인식을 위한 어노테이션항목

〈표 Ⅱ-10〉 텍스트 내 개체명 인식을 위한 어노테이션항목

No	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].NE.id	개체명 식별자	string	선택	LC_0012 (개체명분류_순번)
2	annotations[].NE.text	개체명 텍스트	string	선택	광화문
3	annotations[].NE.type	개체명 타입	string	선택	관광명소 (LC_TOUR)
4	annotations[].NE.begin	객체명 구성 첫 형태소 식별자	string	선택	LC_0009
5	annotations[].NE.end	객체명 구성 끝 형태소 식별자	string	선택	LC_0015
6	annotations[].NE.weight	개체명 인식 결과 신뢰도	number	선택	0.92 (0 ~ 1)

● 관계-의존성 어노테이션 라벨

〈표 Ⅱ-11〉 관계-의존성 어노테이션 라벨

No	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].dependency[].id	어절 ID (출현 순서)	string	선택	DEF_0021 (분류_순번)
2	annotations[].dependency[].text	의존구문 텍스트	string	선택	안녕하세요. 좋은 아침입니다.
3	annotations[].dependency[].head	부모 어절의 ID	string	선택	DEF_0020
4	annotations[].dependency[].label	의존관계 레이블	string	선택	-
5	annotations[].dependency[].mod[]	자식 어절들의 ID	string	선택	-
6	annotations[].dependency[].weight	의존구문 분석 결과 신뢰도	number	선택	0~1

3 OCR 이미지 데이터

3.1 라벨링 공통 항목

● 라벨링 메타정보 공통참조항목

〈표 Ⅱ-12〉 라벨링 메타정보 공통참조항목

No	속성명	항목 설명	Type	필수여부	작성예시
1	Dataset.identifier	데이터셋 식별자	string	필수	IMG_OCR_01 (데이터유형_목적_순번)
2	Dataset.name	데이터셋 이름	string	필수	이미지 내 간판 텍스트 인식을 위한 학습용 데이터셋
3	Dataset.src_path	데이터셋 폴더 위치	string	필수	/dataSet/text/
4	Dataset.label_path	데이터셋 레이블 폴더 위치	string	필수	/dataSet/text/
5	Dataset.category	데이터셋 카테고리	number	필수	0: OCR, 1: 객체인식 등
6	Dataset.type	데이터셋 타입	number	필수	0: 텍스트, 1: 이미지, 2:영상, 3: 음성 등

● 라벨링 이미지 파일 공통참조항목

〈표 Ⅱ-13〉 라벨링 이미지 파일 공통참조항목 - 1

No	속성명	항목 설명	Type	필수여부	작성예시
1	Images.identifier	이미지 식별자 (파일명)	string	필수	IMG_OCR_01_00001 (Dataset ID_순번)
2	Images.type	이미지 파일 확장자	string	필수	JPG, PNG 등
3	Images.width	이미지 가로 크기 (픽셀)	number	필수	1012
4	Images.height	이미지 세로 크기 (픽셀)	number	필수	768
5	data_captured	이미지 생성 일자	string	필수	yyyy.mm.dd HH:MM:SS

※ 데이터 전처리를 통해 이미지 내 텍스트 영역(바운딩박스)만 따로 추출한 경우에는 바운딩박스 정보(x, y, width, height)는 생략 가능

● 저작권 정보가 존재할 때, 선택적으로 작성

〈표 Ⅱ-14〉 라벨링 이미지 파일 공통참조항목 - 2

No	속성명	항목 설명	Type	필수여부	작성예시
1	licenses.id	라이센스 고유 번호	string	선택	http://www.apache.org/licenses/LICENSE-1.0
2	licenses.name	라이센스 이름	string	선택	Apache License 1.0
3	licenses.url	문서 식별자	string	선택	NEWS_000001

3.2 어노테이션 공통 항목

● 이미지 데이터 유형의 라벨링 기능 및 어노테이션 방식

〈표 Ⅱ-15〉 텍스트 데이터 라벨링 기능 및 어노테이션 방식

No	라벨링 기능	어노테이션 방식
1	• 이미지 분류(Image Classification)	• 클래스 라벨(단일, 다중)
2	• 객체 인식(Object Recognition)	• 바운딩 박스(사각형) • 폴리곤(다각형)
3	• 영역 구분(Segmentation)	• 픽셀(점)

● 이미지 내 텍스트 영역에 대한 사각형 박스 형태의 어노테이션 구조

〈표 Ⅱ-16〉 바운딩박스(Bounding Box) 어노테이션 구조

No	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].bbox.id	바운딩박스 식별자	string	필수	BBX_0001 (분류_순번)
2	annotations[].bbox.text	바운딩박스 내 텍스트	string	필수	교보문고
3	annotations[].bbox.x	바운딩박스 시작점 x 좌표	number	필수	100 (좌측상단 기준)
4	annotations[].bbox.y	바운딩박스 시작점 y 좌표	number	필수	120 (좌측상단 기준)
5	annotations[].bbox.width	바운딩박스 가로 길이(픽셀)	number	필수	273
6	annotations[].bbox.height	바운딩박스 세로 길이(픽셀)	number	필수	125

● 이미지 내 텍스트와 매핑되는 실제 장소가 존재할 때, 매핑 관계를 표현하기 위한 선택적 항목

〈표 Ⅱ-17〉 매핑 관계를 표현하기 위한 선택적 항목

No	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].place.id	장소 식별자	string	선택	PLC_0001
2	annotations[].place.title	장소명	string	선택	교보문고 광화문점
3	annotations[].place.addr	장소 주소	string	선택	서울특별시 종로구 종로1가 종로 1
4	annotations[].place.longitude	위치정보(경도)	string	선택	126.977759
5	annotations[].place.latitude	위치정보(위도)	string	선택	37.570975

※ 관심지점(POI, Point of Interest)의 위치 정보는 도로명 주소와 WGS84 좌표 체계 혹은 국가지점번호 체계 활용
※ 공개 제한된 관심지점의 위치 정보는 표시하지 않음

4 자율주행 데이터

4.1 라벨링 공통 항목

- 라벨링 메타정보 공통참조항목

〈표 Ⅱ-18〉 라벨링 메타정보 공통참조항목

No	속성명	항목 설명	Type	필수여부	작성예시
1	Dataset.identifier	데이터셋 식별자	string	필수	AUD_01 (데이터유형_순번)
2	Dataset.name	데이터셋 이름	string	필수	자율 주차를 위한 학습용 데이터셋
3	Dataset.src_path	데이터셋 폴더 위치	string	필수	/dataSet/text/
4	Dataset.label_path	데이터셋 레이블 폴더 위치	string	필수	/dataSet/text/
5	Dataset.category	데이터셋 카테고리	string	필수	0: 운행중데이터, 1: 정지객체인식 등
6	Dataset.type	데이터셋 타입	string	필수	(복수개의 데이터가 존재하므로 빈값으로 설정)

- 라벨링 이미지 파일 공통참조항목

〈표 Ⅱ-19〉 라벨링 이미지 파일 공통참조항목

No	속성명	항목 설명	Type	필수여부	작성예시
1	Images.identifier	이미지 식별자 (파일명)	string	필수	AUD_01_IMG_00001 (Dataset ID_유형_순번)
2	Images.type	이미지 파일 확장자	string	필수	JPG, PNG 등
3	Images.width	이미지 가로 크기(픽셀)	number	필수	1012
4	Images.height	이미지 세로 크기(픽셀)	number	필수	768
5	Images.data_captured	이미지 생성 일자	string	필수	2020-12-29 12:40:23 (yyyy-MM-dd HH:mm:ss)
6	Images.frame_num	영상 내 이미지 프레임 순서	number	필수	573

- 라벨링 영상 파일 공통참조항목

〈표 Ⅱ-20〉 라벨링 영상 파일 공통참조항목 - 1

No	속성명	항목 설명	Type	필수여부	작성예시
1	Images.identifier	영상 식별자 (파일명)	string	필수	AUD_01_MOV_00001 (Dataset ID_유형_순번)
2	Images.type	영상 파일 확장자	string	필수	JPG, PNG 등
3	Images.width	영상 가로 크기 (픽셀)	number	필수	1012
4	Images.height	영상 세로 크기 (픽셀)	number	필수	768
5	Images.data_captured	영상 생성 일자	string	필수	2020-12-29 12:40:23 (yyyy-MM-dd HH:mm:ss)
6	Images.play_time	영상 길이	string	필수	00:19:21 (HH:mm:ss)

- 저작권 정보가 존재할 때, 선택적으로 작성

〈표 Ⅱ-21〉 라벨링 영상 파일 공통참조항목 - 2

No	속성명	항목 설명	Type	필수여부	작성예시
1	licenses.id	라이센스 고유 번호	string	선택	http://www.apache.org/licenses/LICENSE-1.0
2	licenses.name	라이센스 이름	string	선택	Apache License 1.0
3	licenses.url	문서 식별자	string	선택	NEWS_000001

4.2 어노테이션 공통 항목

- 이미지 내 객체 영역에 대한 사각형 박스 형태의 어노테이션 구조

〈표 Ⅱ-22〉 바운딩박스(Bounding Box) 어노테이션 구조

No	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].bbox.id	바운딩박스 식별자	string	선택	BBX_0001 (분류_순번)
2	annotations[].bbox.name	바운딩박스 내 객체명	string	선택	승용차
3	annotations[].bbox.category	바운딩박스 내 객체 유형	string	선택	Car
4	annotations[].bbox.x	바운딩박스 시작점 x 좌표	number	선택	100 (좌측상단 기준)
5	annotations[].bbox.y	바운딩박스 시작점 y 좌표	number	선택	120 (좌측상단 기준)
6	annotations[].bbox.width	바운딩박스 가로 길이(픽셀)	number	선택	273
7	annotations[].bbox.height	바운딩박스 세조 길이(픽셀)	number	선택	125
8	annotations[].bbox.longitude	객체 위치(경도)	string	선택	126.977759
9	annotations[].bbox.latitude	객체 위치(위도)	string	선택	37.570975

● 이미지 내 객체 영역에 대한 3차원 육면체 박스 형태의 어노테이션 구조

〈표 Ⅱ-23〉 3D 바운딩박스(3D Bounding Box, Cuboid) 어노테이션 구조

No	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].3dbbox.id	3D 바운딩박스 식별자	string	선택	CBD_0001 (분류_순번)
2	annotations[].3dbbox.name	3D 바운딩박스 내 객체명	string	선택	승용차
3	annotations[].3dbbox.category	3D 바운딩박스 내 객체 유형	string	선택	Car
4	annotations[].bbox.vertices[]	3D 바운딩박스 꼭지점 좌표	number	선택	[(10, 10, -10), ...] [(x, y, z), ...]
5	annotations[].bbox.edges[]	3D 바운딩박스 꼭지점 2개를 연결하는 변 좌표	number	선택	[(10, 10, 30, 10),...] [(x1, y1, x2, y2), ...]
6	annotations[].bbox.longitude	객체 위치(경도)	string	선택	126.977759
7	annotations[].bbox.latitude	객체 위치(위도)	string	선택	37.570975

● 이미지 내 객체 영역에 대한 다각형 형태의 어노테이션 구조

〈표 Ⅱ-24〉 폴리곤(Polygon) 어노테이션 구조

No	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].polygon.id	폴리곤 식별자	string	선택	PLG_0001 (분류_순번)
2	annotations[].polygon.name	폴리곤 내 객체명	string	선택	승용차
3	annotations[].polygon.category	폴리곤 내 객체 유형	string	선택	Car
4	annotations[].polygon.points[]	폴리곤 내 점(x, y)의 집합	number	선택	[(100, 105), ...,(160,104)]

● 이미지 내 영역에 대한 선 형태의 어노테이션 구조

〈표 Ⅱ-25〉 폴리라인(Polyline) 어노테이션 구조

No	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].polyline.id	폴리라인 식별자	string	선택	PLL_0001 (분류_순번)
2	annotations[].polyline.name	폴리라인 내 객체명	string	선택	중앙선
3	annotations[].polyline.category	폴리라인 내 객체 유형	string	선택	CenterLine
4	annotations[].polyline.points[]	폴리라인 내 점(x, y)의 집합	number	선택	[(100, 105), ...,(160,104)]

● 이미지 내 객체 영역에 대한 다각형 형태의 어노테이션 구조

〈표 Ⅱ-26〉 세그멘테이션(Segmentation) 어노테이션 구조

No	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].segm.id	세그멘테이션 식별자	string	선택	SEG_0001 (분류_순번)
2	annotations[].segm.name	세그멘테이션 내 객체명	string	선택	승용차
3	annotations[].segm.category	세그멘테이션 내 객체 유형	string	선택	Car
4	annotations[].segm.points[]	세그멘테이션 내 점(x, y)의 집합	number	선택	[(100, 105), ...,(160,104)]

- 영상 내 발생 이벤트 어노테이션 구조

〈표 Ⅱ-27〉 이벤트(Event) 어노테이션 구조

No	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].event .id	이벤트 식별자	string	선택	EVT_0001 (분류_순번)
2	annotations[].event .name	이벤트명	string	선택	자동차추돌사고
3	annotations[].event .category	이벤트 유형	string	선택	Accident
4	annotations[].event .start	영상 내 이벤트 시작 시점	number	선택	00:19:21 (HH:mm:ss)
5	annotations[].event .end	영상 내 이벤트 종료 시점	number	선택	00:19:32 (HH:mm:ss)
6	annotations[].event .entities[]	이벤트 관련 엔티티 목록	string	선택	[BBX_0001, BBX_0002] 엔티티 ID 목록

5 영상(동적/정적) 이미지 데이터

5.1 라벨링 공통 항목

- 영상(동적/정적) 이미지 획득 시 라벨링을 수행하기 위한 필수 반영 요소

〈표 Ⅱ-28〉 획득 공통 참조 항목(디지털 카메라, 스마트폰)

No	속성명	항목 설명	작성예시
1	identifier/filename	파일명	세글자 이름(예: “DSC_0001.JPG”)
2	date	촬영날짜(년, 월), 시간	2020.11.20. 17:08:15
3	file format	파일 형식(포맷)	TIFF/JPG/PNG/MP4/AVI 등
4	imsize	이미지 파일 크기	4800KB
5	images_Photographer	촬영자	촬영한 사람
6	device(camera, lidar)	장비정보	스마트폰, 디지털 카메라, drone, CCTV
7	region_name	촬영 지역명	서울시 종로구
8	images_location	촬영위치	강남구 영동대로 스타벅스
9	copyright	저작권 정보	저직권 정보 첨부 필드 체크

No	속성명	항목 설명	작성예시
10	Video Clip	촬영시간	2~6분, 20~40분
11	length	영상길이	5분 영상에서 3분10초 부분 사용
12	FPS/Frame Rate	1초/프레임 재생 속도	30fps
13	width, height	이미지 사이즈	이미지 크기 4031*3024
14	Aspect ratio	비율(종횡비)	16:9(동영상)4:3(이미지)/가로세로
15	resolution	해상도	가로X세로 예) FHD(1920X1080)
16	bit	비트값	컬러색상/기본 24bit
17	Pixel	화소	사진정보/색상정보값(이미지픽셀)
18	depth	RGB 여부	색대표 : RGB, sRGB 등/비트값과 연관
19	ISO	ISO 감도	밝기에 따른 필름 감도
20	definition	선명도	일반 낮음 높음
21	white balance	화이트 밸런스	K(캘빈)단위 색온도/백열등, 형광등
22	exposure time	노출시간	조리개+셔터스피드 값
23	Exposure mode	노출 모드	자동 노출
24	Metering mode	측광 모드	스팟(중앙 중심) 접사에서 주로 사용
25	F-Stop	조리개 값	f2.8~f11 까지 이미지 밝기 조절
26	flash	플래시	자동 / 플래시 터지지 않음
27	filter	필터	필터 여부
28	focal length	초점 거리	mm초점거리/35mm~50mm(표준)
29	FOV(Field of View)	시야각(화각)	35mm → 63도 예)50mm → 46도
30	angle	촬영각도	촬영(360도 회전하며 8가지 이상)
31	GPS(Latitude,Longitude)	GPS 정보(위도, 경도)	GPS/GLONASS/37°30'24.7", 126°53'22.1"
32	weather	날씨정보	1)맑음 2)흐림 3)비 4)눈 중 선택

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

- 영상(동적/정적) 이미지 획득 시 라벨링을 수행하기 위한 필수 반영 요소

〈표 Ⅱ-29〉 획득 선택항목(CCTV)

No	속성명	항목 설명	작성예시
1	Visible distance	가시거리	최소 10m 이내

- 영상(동적/정적) 이미지 획득 시 라벨링을 수행하기 위한 필수 반영 요소

〈표 Ⅱ-30〉 획득 선택 항목(드론/위성)

No	속성명	항목 설명	작성예시
1	Mounted sensor	탑재 센서	1/2.3“유효픽셀수:12M
2	Flight time	비행 시간	23분
3	frequency	송신기 주파수	2.4GHz ISM
4	temperature	온도	-30℃ ~ 220℃
5	humidity	습도	0~100%,정확도
6	coordinates	영상좌상단, 후하단좌표	촬영 고도에 따른 지상기준 점 설정 값
7	INS	카메라 회전각 정보	X,Y,Z 3축 짐벌
8	speed(hoboring,1m/s,2m/s, 4m/s, 8m/s)	비행속도	X(Pitch),Y(Roll),Z(Yaw)방향의 전량
9	Range(m/s)	촬영범위	30~300cm
10	altitude	촬영고도	150m~
11	overlap	중복도	GSD(Ground Sampling Distance) 확보
12	Ascent, Descent speed	최대 상승, 하강 속도	5m/s, 3m/s
13	mission	촬영지 분류	산림지, 관광지, 도심지
14	Working temperature	작동 온도	0℃ ~ 40℃

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

- 영상(동적/정적) 이미지 획득 시 라벨링을 수행하기 위한 필수 반영 요소

〈표 Ⅱ-31〉 특수 카메라

No	항목명	특수 촬영 목적
1	CT, MRI, X-ray 등	헬스케어
2	열화상 카메라	피복, 체온, 상하수도, 시설물 등 주변 온도 변화 (지하 하수구 누수)
3	LiDAR (센서)	자율 주행, 3D 객체 인지, 지형 탐색
4	수중 카메라(CCTV)	수중 영상
5	짐벌, 액티브 캠, 초분광, 적외선	특수 촬영 목적으로 활용

5.2 어노테이션 공통 항목

- 어노테이션 형식 및 정의

〈표 Ⅱ-32〉 어노테이션 형식 및 정의

No	어노테이션 형태	항목 설명
1	annotations[].id	어노테이션 식별자
2	annotations[].image_id	연관 영상(동적/정적) 이미지 식별자
3	annotations[].classes	어노테이션 클래스
4	annotations[].segmentation	객체 영역 정보
5	annotations[].bbox	어노테이션 바운딩박스 정보
6	annotations[].polygon	어노테이션 폴리곤 정보
7	annotations[].polyline	어노테이션 폴리라인 정보
8	annotations[].cuboid	어노테이션 큐보이드 정보
9	annotations[].points	어노테이션 포인트 정보

※ 라벨링 작업 시 예로 바운딩 박스의 시작 좌표와 이어지는 좌표, 끝점 좌표가 매우 중요함

- 공통 참조 필수항목(디지털 카메라, 스마트폰, CCTV, 드론/위성) : 공통 참조기준을 바탕으로 어노테이션 항목으로 변환 (예시로서 활용 가능)

〈표 Ⅱ-33〉 영상(동적/정적) 이미지 데이터 라벨링 정보

No	속성명	항목 설명	Type	필수여부	작성예시
1	videos[].filename	파일 이름	string	필수	DSC_0001 (분류_순번)
2	videos[].id	ID	string	필수	DSC_0001_개체번호 (분류_순번)
3	videos[].date_created	촬영일자	string	필수	2020.11.20. 17:08:15
4	videos[].type	데이터 형식	string	필수	mp4, PNG, JPG
5	videos[].format	포맷	string	필수	h.264/mpeg-4
6	videos[].filesize	크기	number	필수	4800KB
7	videos[].photographer	촬영자	string	필수	홍길동
8	videos[].device	촬영 장비	string	필수	디지털 카메라
9	videos[].location	촬영 지역명	string	필수	서울시 종로구 (동까지만 표기)

No	속성명	항목 설명	Type	필수여부	작성예시
10	videos[].license	라이선스	string	필수	-
11	videos[].length	영상길이	string	필수	10M
12	videos[].FPS	프레임 재생속도	string	필수	30
13	videos[].frames	총 프레임 수(FPS)	number	필수	60
14	videos[].aspect_ratio	종횡비	string	필수	4:3
15	videos[].width	너비	number	필수	4031
16	videos[].height	높이	number	필수	3024
17	videos[].resolution	해상도	string	필수	FHD
18	videos[].bit	비트값	string	필수	24bit
19	videos[].pixel	화소	string	필수	4K
20	videos[].color_depth	색심도	string	필수	sRGB
21	videos[].ISO	ISO 감도	string	필수	3200
22	videos[].whith balance	화이트 밸런스	string	필수	5500K
23	videos[].exposure_time	노출시간	string	필수	f2.8 1/80
24	videos[].F-stop	조리개값	string	필수	f2.8
25	videos[].flash	플래시	string	필수	자동
26	videos[].focal_length	초점거리	string	필수	50mm
27	videos[].angle_view	화각	string	필수	46
28	videos[].angle	촬영각도	string	필수	120도
29	videos[].weather	날씨정보	string	필수	맑음

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

- cctv 영상(동적/정적) 이미지 - 공통 참조기준을 바탕으로 어노테이션 항목으로 변환 (예시로서 활용 가능)

〈표 Ⅱ-34〉 장비별 영상(동적/정적) 이미지 참조 항목 라벨링 정보 - 1

No	속성명	항목 설명	Type	필수여부	작성예시
1	videos[].visible_distance	가시거리	number	필수	50M
2	videos[].temperature	온도	number	선택	3℃
3	videos[].humidity	습도	number	선택	32%
4	videos[].coordinates	좌표	number	선택	위경도 정보
5	videos[].cctv_name	CCTV 명	string	필수	광화문4거리 3번CCTV

No	속성명	항목 설명	Type	필수여부	작성예시
6	videos[].range	촬영범위(360도 회전 혹은 고정)	string	필수	50
7	videos[].mission	촬영지 분류	string	필수	도심지
8	videos[].event_id	이벤트 분류	string	선택	ABA_0001 (분류_순번)
9	videos[].event_name	이벤트명	string	선택	특이 상황판별
10	videos[].event_name.start_time	이벤트 시작시간	string	선택	2020.11.20. 17:08:15
11	videos[].event_name.end_time	이벤트 종료시간	string	선택	2020.11.20. 17:08:20

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

- 드론/위성 영상(동적/정적) 이미지

〈표 Ⅱ-35〉 장비별 영상(동적/정적) 이미지 참조 항목 라벨링 정보 - 2

No	속성명	항목 설명	Type	필수여부	작성예시
1	videos[].drone_name	드론명	string	필수	DRN_0001 (분류_순번)
2	videos[].sensor	센서	string	선택	가속도
3	videos[].max_flight	최대비행시간	string	선택	1H
4	videos[].temperature	온도	number	선택	3℃
5	videos[].humidity	습도	number	선택	30%
6	videos[].coordinates	좌표	number	필수	위경도 정보
7	videos[].GPS	GPS 정보(위도, 경도)	string	필수	37.3024, 126.53221
8	videos[].INS	카메라 회전각 정보	string	선택	120, 60, 0
9	videos[].speed	비행속도	number	선택	8km/h
10	videos[].range	촬영범위(360도 혹은 고정, 가시거리), 정사영상 등	string	필수	50M
11	videos[].altitude	촬영고도	number	필수	150M
12	videos[].mission	촬영지 분류	string	필수	산림지
13	videos[].overlap	중복도	string	선택	3M, 3M
14	videos[].GCP	지상기준점(GPS와는 별개로 위치 보정 역할 지정)	string	선택	위경도 좌표

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

Ⅲ
데이터 구축사례 모음

제1장 개요

1 추진 배경 및 목적

- 과학기술정보통신부(이하 '과기부')와 한국지능정보사회진흥원(이하 '지능정보원')은 「지능정보산업 인프라 조성」의 일환으로 2017년부터 매년 '인공지능 학습용 데이터 구축 사업'을 추진하며, 381종의 인공지능 학습용 데이터를 구축하여 민간에 개방하는 등 가시적 성과를 확보하고 있다. '인공지능 학습용 데이터 구축 사업'은 '디지털 뉴딜'의 핵심 프로젝트인 '데이터 댐'의 대표 사업으로 인공지능 서비스 개발에 필수적인 양질의 학습용 데이터를 대규모로 구축하고 민간에 개방함으로써 국내 인공지능 산업 생태계 활성화의 마중물 역할이 되는 사업이다.
- 반면, 개방된 데이터를 활용하는 몇몇 수요처 중에서는 인공지능 학습용 데이터의 품질에 대한 이슈를 제기한 바 있다. 또한, 인공지능 학습용 데이터 구축 사업 확대에 따라 수행기관의 데이터 구축 역량과 노하우 차이로 인하여 데이터 품질의 편차가 발생하고 있다. 따라서 구축 데이터 간의 품질 수준의 편차를 최소화하기 위하여 본 『데이터 구축사례 모음』(이하, '『구축사례 모음』')을 마련하였다. 이를 통하여 데이터 구축 초기 단계부터 최종 검증 단계까지 양질의 품질관리 체계 확보에 도움이 되고자 한다.
- 본 『구축사례 모음』은 2020년에 구축된 데이터를 기준으로 14종의 데이터를 임의로 선정하였으며, 해당 수행기관의 데이터 구축 노하우를 살펴볼 수 있다. 14종의 데이터를 비롯한 공개 데이터의 보다 상세한 내용은 'AI 허브'에서 확인 할 수 있다.
- '인공지능 학습용 데이터 구축 사업'에 참여하는 수행기관은 데이터에 대한 이해도와 임무 정의를 명확히 하기 위하여 품질관리체계를 확립하는데 기반이 되는 문서(데이터 구축 및 활용 계획서 등)를 작성하게 된다. 본 『구축사례 모음』은 중소벤처, 연구자, 학생 등 누구나 인공지능 학습용 데이터를 구축하고 활용할 수 있도록 '데이터 구축 및 활용 계획서' 등 관련 문서를 작성하는 데 도움이 되고자 한다.

2 데이터 구축사례 모음 구성 및 활용

- 본 『구축사례 모음』은 아래의 표와 같이 14종의 데이터를 대상으로, 각 수행기관이 작성한 '구축·활용 가이드라인'을 재구성하여 소개한다.

〈표 Ⅲ-1〉 20년도 분야별 데이터 구축사례 데이터

구분	데이터 분류			데이터명
	대분류	중분류	소분류	
1	이미지	비전	JPEG	고서 한자 인식(OCR) AI데이터
2	오디오	자연어	WAV	자유대화 AI 데이터
3	오디오	자연어	WAV	한국어 방언 AI 데이터
4	텍스트	자연어	SNS	한국어 SNS 데이터
5	이미지	비전	JPG, PNG, MP4	K-POP 안무영상 데이터
6	이미지	비전	PNG	고해상도 LF AI 학습용 데이터
7	이미지	헬스케어	X-ray, CT, PET-CT	폐암 AI 학습 데이터
8	이미지	헬스케어	X-ray, CT, PET-CT	갑상선암 AI 학습 데이터
9	이미지	헬스케어	DICOM, MRI	유방암 AI 학습데이터
10	이미지	비전	JPG	주행환경 정적객체 데이터
11	이미지	비전	JPEG	시설작물 질병진단 이미지 데이터
12	이미지	비전	JPG	동의보감 약초 이미지 AI 데이터
13	이미지	비전	JPG	CCTV 영상 AI 데이터
14	비디오	비전	MP4	패션상품 및 착용 영상 AI 데이터

- 각 데이터는 구축활용가이드 절차에 따라 제 1장부터 제 5장까지 구성되며, 세부 절차는 아래 표와 같다.

〈표 Ⅲ-2〉 데이터별 구축활용가이드 절차 및 내용

구축활용가이드 절차	세부 절차
제 1장. 데이터 정보 요약	1. 가이드 분류
	2. 데이터 정보
	3. 데이터 구축 개요
	4. 구축 목적
	5. 활용 분야
	6. 유의 사항

구축활용가이드 절차	세부 절차
재 2장. 데이터 획득 및 정제	1. 원시 데이터 선정
	2. 규제관련 사항
	3. 획득 및 정제 절차
	4. 획득 및 정제 기준
제 3장. 어노테이션/라벨링	1. 어노테이션/라벨링 절차
	2. 어노테이션/라벨링 기준
	3. 어노테이션/라벨링 교육
	4. 어노테이션/라벨링 도구 및 사용법
제 4장. 데이터 검수	1. 검수 절차
	2. 검수 기준
제 5장. 데이터 활용 방안	1. 학습 모델
	2. 서비스 활용 시나리오

● **〈제 1장. 데이터 정보 요약〉**

- (개요) 데이터 구축 초기 단계에서 데이터에 대한 이해도를 높이고 데이터 구축 임무정의를 하는 단계다. 데이터 분류, 데이터 정보, 데이터 구축 개요, 데이터 구축 목적, 데이터 활용 분야, 데이터 구축 시 유의 사항을 작성했다.
- (가이드 분류) 데이터를 대분류*, 중분류**, 소분류***로 구분하여 작성했다. 여기서 의미하는 데이터는 원천데이터임을 유의한다.

  *(대분류) 텍스트, 비디오, 이미지, 오디오
  **(중분류) 음성·자연어, 비전, 헬스케어, 교통·물류, 농축수산, 재난·안전·환경
  ***(소분류) JPEG, WAV, MP4, TXT 등
- (데이터 정보) 데이터명, 활용분야, 데이터요약, 데이터 출처 등을 작성했다.
- (데이터 구축 개요) 데이터 구축 절차에 대하여 전반적인 절차로 데이터 수집, 정제, 가공, 검수 등의 단계에 대한 내용을 설명한다.
- (구축 목적) 데이터 구축 목적을 관련 산업 분야 및 연구 분야와 연관 지어 설명한다.
- (활용 분야) 구축한 인공지능 학습용 데이터셋으로 활용 가능한 분야를 설명한다.
- (유의 사항) 본 데이터 구축 시, 발생 가능한 다양한 문제점을 예방하기 위한 유의사항(데이터 처리, 규제, 저작권, 개인정보처리 등)에 대한 설명을 한다.

- **〈제 2장. 데이터 획득 및 정제〉**
  - (개요) 데이터 수집부터 시작하여 데이터를 라벨링 공정에 투입하기 위해 필요한 전처리 등의 정제까지의 과정을 설명한다.
  - (원시 데이터 선정) 원시 데이터 종류, 선정 방법 및 기준, 수집 량 등 수집하려는 원시 데이터에 관하여 설명한다.
  - (규제관련 사항) 데이터 수집 시, 발생 가능한 문제점을 예방하기 위해 관련 규제 사항에 대하여 설명한다.
  - (획득 및 정제 절차) 수집하려는 데이터의 획득 및 정제 절차에 대한 설명을 한다.
  - (획득 및 정제 기준) 데이터 획득 및 정제 시, 고려해야할 기준을 설명한다.
- **〈제 3장. 어노테이션/라벨링〉**
  - (개요) 어노테이션/라벨링 공정에 대하여 전반적으로 설명한다. 어노테이션/라벨링 방법, 절차, 기준, 작업자(검수자) 교육, 어노테이션/라벨링 도구 및 사용법에 관하여 설명한다.
  - (어노테이션/라벨링 절차) 어노테이션 종류와 라벨링 방법을 설명한다.
  - (어노테이션/라벨링 기준) 어노테이션 및 라벨링 시, 고려해야할 기준을 설명
  - (어노테이션/라벨링 교육) 작업자(검수자)에게 어노테이션/라벨링 관련 필요한 사항에 대한 교육과 교육 방법 등을 설명한다.
  - (어노테이션/라벨링 도구 및 사용법) 어노테이션/라벨링에 활용한 저작도구를 소개하고, 저작도구 사용방법을 설명한다.
- **〈제 4장. 데이터 검수〉**
  - (개요) 어노테이션/라벨링 공정 과정을 거친 데이터에 대한 검수 절차와 기준을 설명한다.
  - (검수 절차) 어노테이션/라벨링된 데이터의 검수 절차를 설명한다.
  - (검수 기준) 어노테이션/라벨링된 데이터의 검수 기준을 설명한다.
- **〈제5장. 데이터 활용 방안〉** 구축한 데이터를 학습하기 위한 인공지능 학습 모델과 서비스 활용 시나리오를 설명한다.
  - (학습 모델) 본 데이터셋을 학습하기 위한 인공지능 학습 모델을 소개하고 선정 근거 등을 설명한다.

- (서비스 활용 시나리오) 구축한 데이터셋과 학습모델을 활용할 수 있는 예시를 설명한다.

〈유의 사항〉

※ 본 『구축사례 모음』의 14개 데이터에 대한 구축활용가이드 절차는 '구축활용가이드(출처: aihub.or.kr)'를 기반으로 작성되었기 때문에, 세부 절차에서 요구하는 내용이 없는 데이터는 세부 절차 내용에 '관련 내용 없음'으로 표기하였음

※ 본 『구축사례 모음』은 분량을 고려하여 내용 중 일부를 생략한 경우도 있으니 보다 상세한 내용은 'AI 허브'에서 원본(구축활용가이드) 참조 요망

제2장 고서 한자 인식(OCR) AI데이터

1 데이터 정보 요약

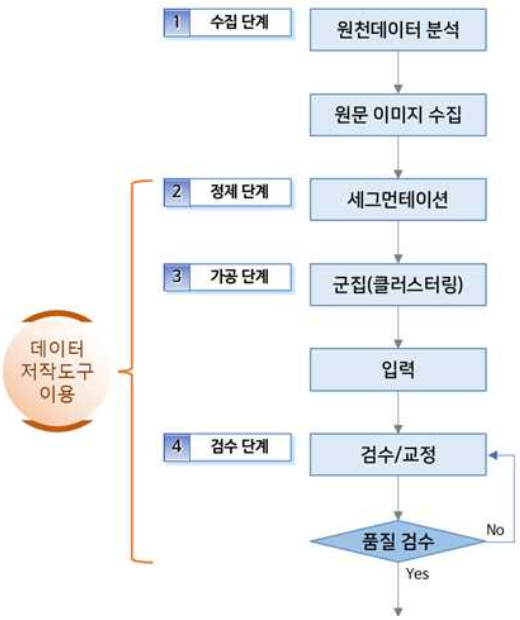
1.1 가이드 분류

대분류	이미지	중분류	비전	소분류	JPEG

1.2 데이터 정보

데이터 이름	고서 한자 인식(OCR) AI데이터
활용 분야	고서 한자 인식(OCR)
데이터 요약	한자로 기록된 국가기록유산(고서, 고문헌 등)의 활용성과 접근성 향상을 위해 고서 이미지 속 한자의 디지털 텍스트를 자동으로 확보하기 위한 인공지능 기반 OCR 기술 개발용 학습 데이터

1.3 데이터 구축 개요



[그림 Ⅲ-1] 학습용 데이터 구축 공정도

〈표 Ⅲ-3〉 학습 데이터 단계별 구축 내용

구축단계	세부공정	내용
수집 단계 (획득 단계)	원천데이터 분석	• 원천데이터 : 고서(고도서) 원문 이미지 • 구축 대상 서체 여부 확인 : 해서체(95%), 행서체(5%) • 이미지 품질 확인 : 해상도, 기울기, 훼손 여부 등
	원천데이터 수집	• 개인정보 비식별화 불필요 : 일체의 개인정보 불포함 • 저작권, 특허권, 초상권 부존재 • 대상 자료 전량은 참여기관인 한국국학진흥원 소장 데이터
정제 단계	세그먼테이션	• 원문이미지 상의 한자를 낱자별로 추출 • 데이터 저작도구 사용, 자동/수동 병행
가공 단계	군집 (클러스터링)	• 한자 세그먼트들을 동일 한자끼리 하나의 그룹으로 클러스터링 • 데이터 저작도구 사용, 자동/수동 병행 • 크라우드 소싱
	입력	• 한자 전문인력들이 군집별 입력 • 데이터 저작도구 사용, 전량 수동 • 크라우드 소싱
검수 단계	검수/교정	• 한자 전문인력들이 입력 완료된 한자를 원문이미지와 대조 검수/교정 • 데이터 저작도구 사용, 전량 수동 • 크라우드 소싱
	품질검수	• 한자 검수 전문인력들이 샘플링(5%) 검수 • 데이터 저작도구 사용, 전량 수동

1.4 구축 목적

- AI모델이 고서 이미지 속의 한자를 자동으로 인식하도록 훈련시키는데 필요한 데이터 구축
- 고서 한자 이미지와 이미지 속의 한자 낱자별 한자 유니코드가 라벨링된 데이터쌍 구축

1.5 활용 분야

- 한자(고문헌) 디지털 텍스트라는 1차 콘텐츠의 효율적 확보를 가능하게 하는 기술기반(인공지능 학습데이터와 기초 모델)의 공유를 통해 연결·확장되는 관련 솔루션과 서비스들의 개발을 촉진시킴으로써 정치, 사회, 역사, 지리, 문화, 예술, 과학, 기술, 사상, 종교 등 다양한 분야에서의 파생 콘텐츠 생산을 유도하여 경제적 부가가치 증대와 관련 산업 일자리 창출]

- 인공지능 기반 한자 글자체 인식(OCR) 기술을 통한 속도감 있는 한자(고문헌) 디지털 텍스트 확보로 동아시아 한자문화권(한국, 중국, 일본, 베트남)에서의 학술·문화·전통 콘텐츠 분야 경쟁력 강화 및 한반도 주변 강대국들의 역사 왜곡(중국의 동북공정, 일본의 독도영유권 주장) 시도에 대한 대응연구 활성화

1.6 유의 사항

- 한자에는 문자의 형태는 같으나 여러 개의 음과 뜻을 가진 '동형이음(의)자'가 존재하며, 동형이음(의)자는 입력자가 읽은 음대로 일단 입력한 후, 후처리를 통하여 각각 통일된 하나의 한자로의 변환작업 필요

  ※ 동형이음(의)자는 육안으로 식별되는 디지털 한자 텍스트의 형태까지는 같으나 유니코드 값이 다르기 때문에 하나의 낱자 한자 이미지는 통일된 하나의 유니코드 한자로 학습데이터를 구축

〈표 Ⅲ-4〉 동형이음 한자 학습데이터 구축 기준 예시

〈예시〉

한자 이미지	뜻/음	디지털 한자 텍스트	유니코드 값	통일하여 구축할 값
金	성 김	金	U+91D1	
	쇠 금	金	U+F90A	√
樂	풍류 악	樂	U+6A02	√
	즐길 락	樂	U+F95C	
	좋아할 요	樂	U+F9BF	

2 데이터 획득 및 정제

2.1 원시 데이터 선정

- 본 과제의 원천데이터는 고문헌 중 약 98% 비율을 차지하는 고서(고도서)를 활용
- 본 과제의 원천데이터인 고서를 구성하는 서체는 해서체 약 95%, 행서체 약 5%로 구성(해서체 약 950만자, 행서체 약 50만자)
- 원천자료인 조선시대 고서는 개인정보, 민감정보, 지적재산권(저작권) 침해정보 불포함

■ 고문헌 자료유형 및 서체 비율 조사결과

• 한국국학진흥원 소장(2018) 고문헌 303,088종(점)(고서 63,848종, 고문서 239,240점)의 글자수 기준으로 고도서가 약 30억 6천만자(98%), 고문서가 약 7천 2백만자(2%)

• 전체 고문헌(고도서+고문서) 중 해서체로 제작된 비율이 83.05%로 추정되며, 해서체와 유사한 행서체까지 포함한다면 비율은 96.71%가 됨. 따라서 고문헌의 대부분은 해서체(+행서체)로 제작된 것으로 추론 가능

• 한국국학진흥원 소장 고문헌(고도서/고문서) 수량 및 서체별 비율(2018년 기준)

서체	고도서		고문서		합계		서체 비율
	종수	글자수	종수	글자수	종수	글자수	
해서	54,250	2,604,000,000	2,384	715,200	56,634	2,604,715,200	83.046%
해행서	259	12,432,000	461	138,300	720	12,570,300	0.401%
행서	8,465	406,320,000	31,930	9,579,000	40,395	415,899,000	13.260%
행초서	792	38,016,000	732	219,600	1,524	38,235,600	1.219%
초서	67	3,216,000	203,733	61,119,900	203,800	64,335,900	2.051%
전서	14	672,000	-	-	14	672,000	0.021%
예서	1	48,000	-	-	1	48,000	0.002%
계	63,848	3,064,704,000	239,240	71,772,000	303,088	3,136,476,000	100.000%

※ 고도서는 평균 1면에 200자의 한자, 1책이 80면~100면, 1종이 3-4책으로 구성되므로 보수적으로 종별 3책, 책당 80면, 면당 200자를 기준으로 글자수 추정.

※ 고문서는 대부분 1면이 1종이며, 대체로 필사 형태이므로 글자수는 내용에 따라 편차가 심하여 종별 평균 글자수를 구하기 어려우나 간찰(편지) 등 글자수가 많은 유형을 기준으로 종별 300자를 평균으로 하여 글자수를 추정

2.2 규제관련 사항

- 공개되는 학습용 데이터는 원시데이터에 대한 의존성 없음
- 개인정보, 민감정보, 지적재산권(저작권) 침해정보 불포함

2.3 획득 및 정제 절차

1) 고서의 책별 디지털 이미지 확인: 이미지 포맷이 JPEG인 책 선별

2) 책을 구성하는 이미지의 해상도 확인: 150dbi 이상의 이미지로 구성된 책 선별

3) 책의 한자 서체를 확인: 해서체와 행서체로 제작된 고서 선별

4) 책의 메타 정보 작성: 서명, 저자명, 문체, 서체, 책별 이미지 면수 작성

2.4 획득 및 정제 기준

- 이미지가 포함된 한자가 잘리거나, 초점이 맞지 않아 문자 자체의 식별이 불가능한 이미지 제외

  ※ 이미지의 노이즈 제거 과정은 실시하지 않음

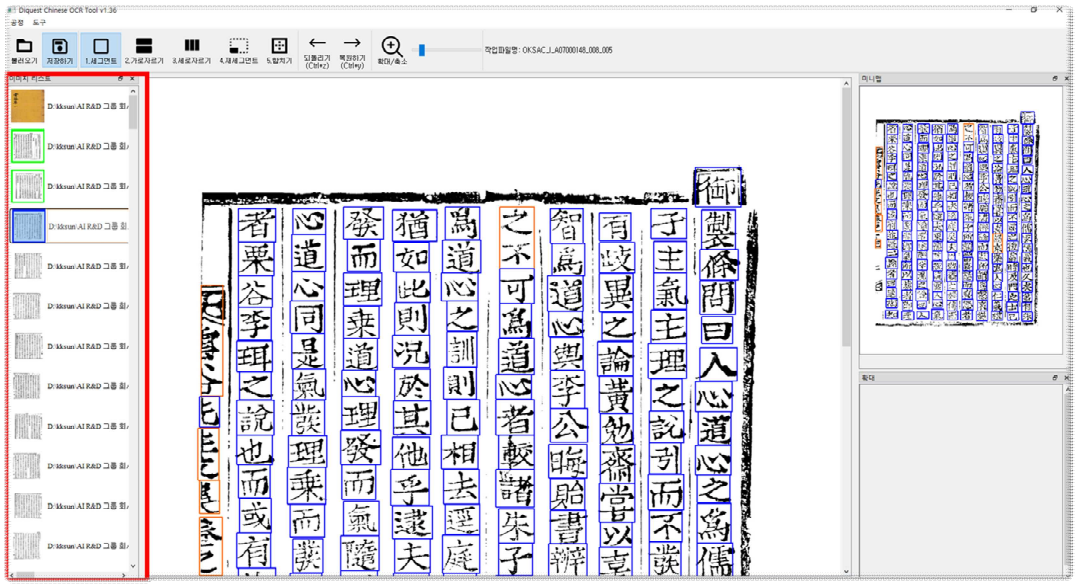
3 어노테이션/라벨링

3.1 어노테이션 / 라벨링 절차

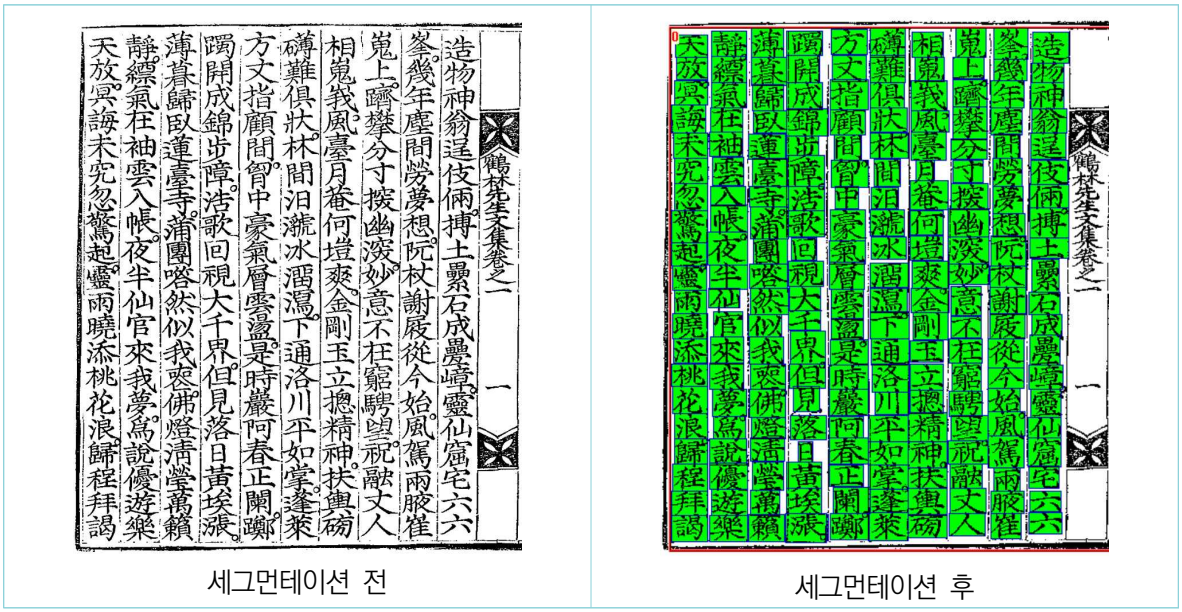
1) 세그먼테이션

- 한자가 포함된 이미지를 화면에 열어 이미지 속의 한자에 한 글자씩 바운딩박스 수행하는 절차
- 데이터 저작도구(DCCIT)가 우선 자동으로 세그먼트를 생성해주며, 사용자가 자동 생성된 세그먼트를 수정하는 방법으로 작업 진행
- 하나의 세그먼트 내에는 한 글자의 한자만 포함되도록 바운딩박스 수행
- 최대한 다른 한자의 획 일부가 세그먼트 대상 한자의 바운딩박스에 포함되지 않도록 바운딩박스 수행
- 원천자료 이미지 상 한자들의 자간이 좁아 한자의 획들이 정사격형의 문자 영역을 상호 침범하는 경우에는 불가피하게 세그먼트 대상 한자의 획 일부가 잘리거나, 다른 한자의 획 일부가 세그먼트 대상 한자의 바운딩박스에 포함 가능. (단, 세그먼트 대상 한자의 획 일부가 잘리거나 다른 한자의 획 일부가 세그먼트에 포함되었을 때 세그먼트 대상 한자의 판독(인식)에 지장을 주거나 다른 한자로 오인되지 않아야 함)

※ 원천자료 이미지 상 한자들의 자간이 좁아 각 세그먼트(바운딩박스) 간의 간섭이 심한 경우 마우스 조작이 능숙한 한자 판독 숙련자가 작업하는 것이 유리함



[그림 Ⅲ-2] 세그먼테이션을 위한 저작도구



세그먼테이션 전　　세그먼테이션 후

[그림 Ⅲ-3] 세그먼테이션 작업 결과

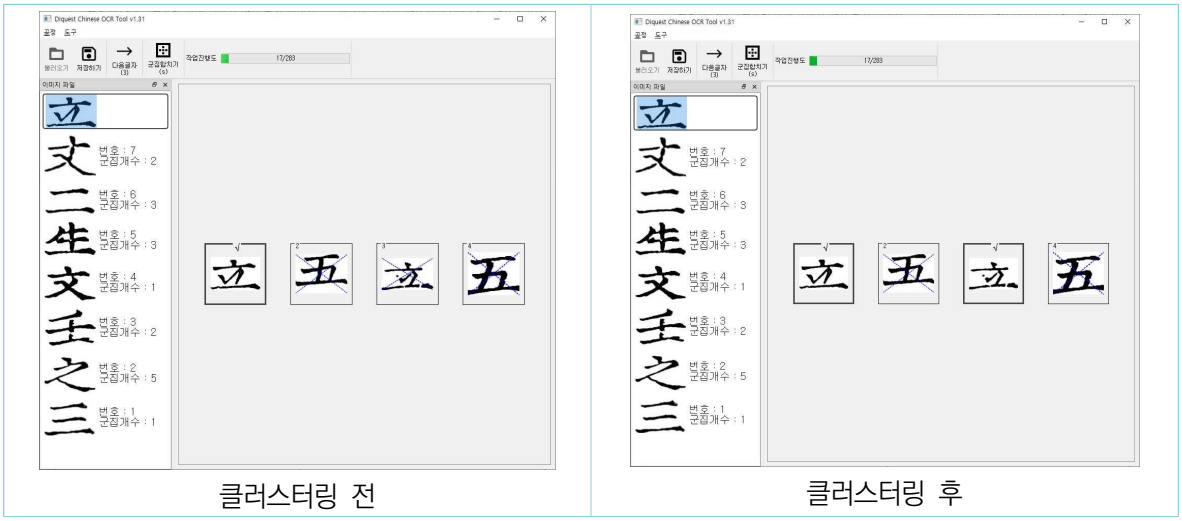
2) 클러스터링

- 세그먼트가 끝난 데이터를 열어 같은 한자 세그먼트를 하나의 그룹(클러스터)으로 묶어주는 절차
- 데이터 저작도구(DCCIT)가 우선 자동으로 클러스터를 생성해주며, 사용자가 자동 생성된 클러스터를 수정하는 방법으로 작업을 진행함

- 불러온 하나의 클러스터 안에 포함된 한자 낱자 이미지들을 확인하여 대표 한자 이미지와 다른 한자가 포함되어있을 경우 해당 클러스터에서 제거함
- 불러온 하나의 클러스터 안에 포함된 한자 낱자 이미지들을 확인하여 대표 한자 이미지와 같은 한자인데도 클러스터에서 제외한다는 표시가 되어있는 한자가 있을 경우 해당 클러스터에 포함해줌
- 대표한자가 동일한 복수의 클러스터가 존재할 수 있으며, 이 때는 가능한 한 하나의 클러스터로 묶어줌

※ 형태상 유사한 한자들의 클러스터 판독 및 수정은 한자 판독 숙련자가 작업하는 것이 검증 및 수정 속도와 품질 확보에 유리함

[그림 Ⅲ-4] 클러스터링을 위한 저작도구



클러스터링 전　　클러스터링 후

[그림 Ⅲ-5] 클러스터링 작업 결과

3) 입력

- 클러스터링이 끝난 데이터를 열어 불러온 클러스터별로 한자 텍스트를 입력해주는 절차
- 대표 한자의 이미지를 확인하여 해당 한자 텍스트 입력

  ※ 유니코드에 반영되어 있지 않아 입력이 불가능한 한자(신출한자, 고한자?)나 문자로 인식은 되지만 원본의 훼손, 마멸 등으로 판독이 불가능하여 역시 입력이 불가능한 한자는 약속된 특수기호 ▼로 입력
- 동형이음(의)자는 통일된 하나의 한자로 입력해야 함(동형이음(의)자 입력 기준 별도 수립)

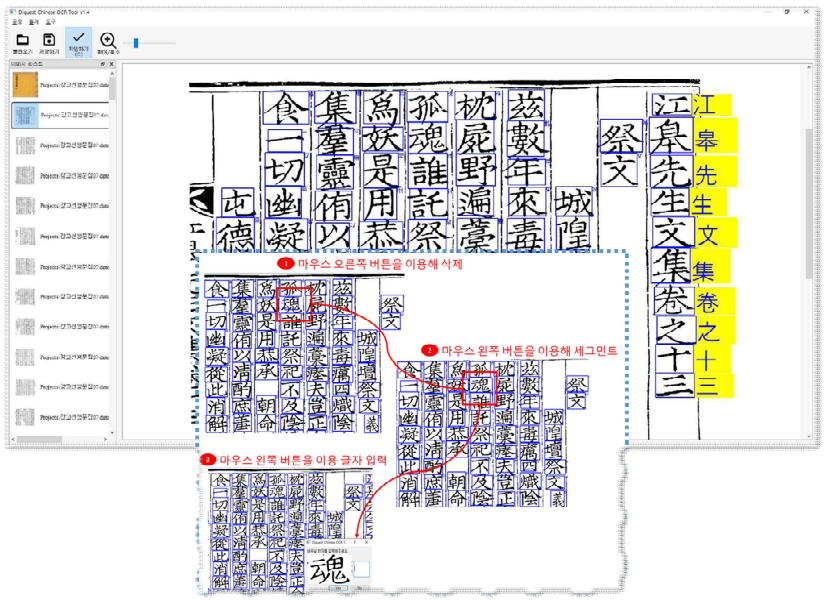
  ※ 한자 판독 및 입력 숙련자가 작업해야 함

[그림 Ⅲ-6] 입력을 위한 저작도구

4) 검수/교정

- 원천자료 이미지와 입력된 한자 텍스트를 낱자별로 대조해가며 정확하게 입력되었는지 비교 확인
- 잘못 입력된 한자는 바로 수정 입력
- 동형이음(의)자는 통일된 하나의 한자로 입력해야 함

  ※ 동형이음(의)자 입력 기준 별도 수립)



[그림 Ⅲ-7] 저작도구를 이용한 검수/교정

3.2 어노테이션 / 라벨링 기준

- 세그먼테이션

  - 하나의 세그먼트 내에는 한 글자의 한자만 포함되어야 함

- 클러스터링

  - 하나의 클러스터에는 동일한 한자만 포함되어야 함

- 입력

  - 한자 이미지를 판독하여 해당하는 정확한 한자를 입력해야 함

- 동형이음(의)자는 통일된 하나의 한자로 입력해야 함(동형이음(의)자 입력 기준 별도 수립)

● 검수/교정

- 한자 낱자 이미지와 정확하게 일치하는 한자 텍스트를 입력해야 함
- 동형이음(의)자는 통일된 하나의 한자로 입력해야 함(동형이음(의)자 입력 기준 별도 수립)

3.3 어노테이션 / 라벨링 교육

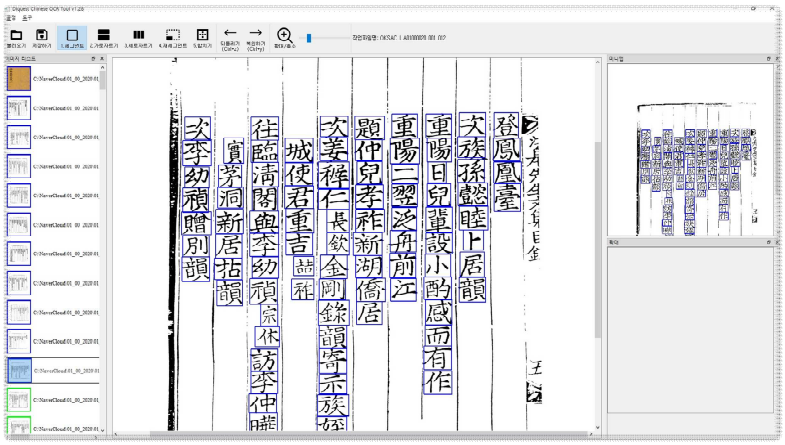
● 획득/정제 작업을 수행하는 인원은 한자 서체에 대한 이해를 보유한 자로서 육안으로 한자의 서체(해서체, 행서체, 초서체 등)를 구분할 수 있어야 함

※ 한자 서체에 대한 이해는 한자 이미지 자료를 사용한 교육훈련으로 습득 가능함

3.4 어노테이션 / 라벨링 도구 및 사용법

● 학습용 데이터 저작도구 (DCCIT : Diquest Chinese Character Input Tool)

- DCCIT는 한자가 포함된 이미지 상에서 한자를 낱자별로 표시(세그먼트)하고 해당 세그먼트별로 유니코드 한자를 입력할 수 있는 도구
- 한자 낱자 기준의 학습 데이터 구축을 위해 한자가 포함된 이미지 상에서 한자 낱글자별로 세그먼트를 생성(바운딩박스 치는 작업)할 수 있는 효율적인 방법(세그먼테이션)을 제공
- 단시간에 많은 한자를 입력할 수 있도록 세그먼테이션이 완료된 일정수량의 한자 이미지들에 대해서 동일한 한자 세그먼트를 하나의 그룹(클러스터)으로 묶을(클러스터링) 수 있는 효율적인 방법을 제공
- 생성된 클러스터별로 유니코드 한자를 입력할 수 있는 한자 입력 기능을 제공. 클러스터에 한자가 한 번 입력되면 해당 클러스터에 포함된 모든 동일 한자 세그먼트(바운딩박스)에 대해서 입력된 한자가 맵핑되어 한꺼번에 입력됨
- 모든 클러스터에 유니코드 한자 입력이 완료되면 원천자료 이미지에 포함된 모든 한자에 대해서 입력이 완료됨
- 입력이 완료된 원천자료 한자 이미지 한 면과 입력된 유니코드 한자 텍스트들을 1:1로 대조해가며 검수 및 교정할 수 있는 효율적인 기능을 제공



[그림 Ⅲ-8] 세그먼테이션 기능

[그림 Ⅲ-9] 클러스터링 기능

[그림 Ⅲ-10] 클러스터링 기능

4 데이터 검수

4.1 검수 절차

1) 품질검수는 구축 및 공개 주데이터인 고서 한자 인식 학습데이터 1천만 건에 대하여 ①세그먼테이션 ②클러스터링 ③입력 ④검수/교정의 4단계 공정을 모두 마친 후 실시함

※ 각 공정 단계에서는 전 단계의 모든 데이터를 활용하게 되므로 자연적으로 중간검수 과정을 거치게 됨

- ②클러스터링 단계에서는 ①세그먼테이션 단계에서 생성된 모든 세그먼트를 사용하므로 모든 세그먼트에 대한 검사효과를 가짐
- ③입력 단계에서는 ②클러스터링 단계에서 생성된 모든 군집에 대한 입력을 진행하므로 모든 군집에 대한 검사효과를 가짐
- ④검수/교정 단계는 ③입력 단계에서 입력된 모든 한자에 대해 진행하므로 모든 입력에 대한 검수효과를 가짐

2) 품질검수 수량 : 구축 및 공개 목표량인 1천만 건(자)의 5% 이상 ※:50만 건(자) 이상

3) 품질검수 대상 선정 : 구축 완료 후 집계된 책별 글자수를 기준으로 글자수 합계가 50만자 이상이 되도록 책 단위로 무작위 샘플링 실시 (클러스터링 데이터의 정확도 검사를 위한 샘플링은 공정관리 단위로 실시

※ 클러스터링은 공정관리 단위로 실시하기 때문에 책 단위로의 분리가 불가능)

4) 품질검수 절차

- 검수 대상 샘플링(책 단위 및 공정관리 단위)
- 세그먼테이션 정확도 검사: 샘플링된 책들의 전체 세그먼트 내 오류 세그먼트 포함 여부를 육안 식별하여 집계
- 세그먼테이션 유효성 검사: 샘플링된 책들의 전체 세그먼트에 대한 기계적 검사
- 클러스터링 정확도 검사: 샘플링된 전체 클러스터에 내 오류 세그먼트 포함 여부를 육안 식별하여 집계함
- 문자인식 정확도 검사: 샘플링된 책들의 전체 세그먼트 내 오입력 한자 포함 여부를 육안 식별하여 집계

4.2 검수 기준

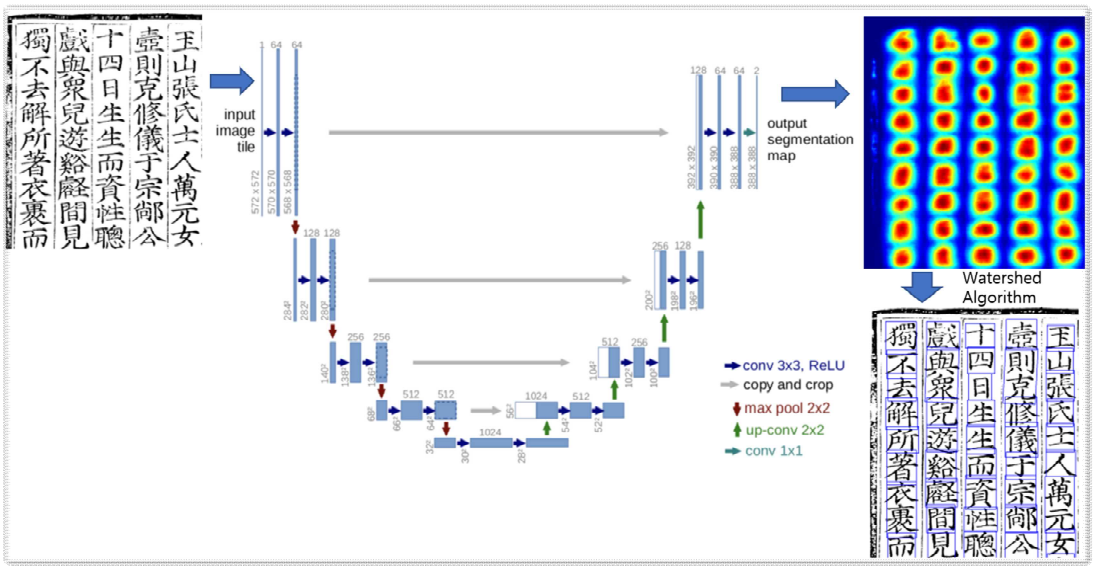
〈표 Ⅲ-5〉 어노테이션 / 라벨링 검수 기준

데이터	구분		측정지표	검수기준	자동화	검사율	샘플링 단위
문자인식 데이터 (주데이터)	정확도	구조 및 형식	• 문자입력 정확도 • 단위 : 이미지 상에서 세그먼테이션된 한자 낱자 1개와 입력된 디지털 텍스트 1자의 쌍 • 방법 : 자료이미지 상의 총 문자수 대비 정확하게 입력된 문자수의 비율 측정	• 정확도 99.9% 이상	수동 (육안)	5% 샘플링 (50만자 이상)	책
세그먼테이션 데이터 (보조 데이터)	정확도	구조 및 형식	• 세그먼테이션 정확도 • 단위 : 이미지 상에서 세그먼테이션된 한자 낱자 1개 • 방법 : 자료이미지 상의 총 문자수 대비 정확하게 세그먼트된 문자수의 비율 측정	• 정확도 99% 이상	수동 (육안)	5% 샘플링 (50만자 이상)	책
	유효성	한자 객체 검출 정확도	• mAP	• mAP 0.64 이상 (Precision 80%, Recall 80% 이상)	자동	5% 샘플링 (50만자 이상)	책
클러스터링 데이터 (보조 데이터)	정확도	구조 및 형식	• 클러스터링 정확도 • 방법 : 자료이미지 상의 총 문자수 대비 정확하게 클러스터링 된 문자수의 비율 측정	• 정확도 99% 이상	수동 (육안)	5% 샘플링 (50만자 이상)	클러스터 생성 단위

5 데이터 활용 방안

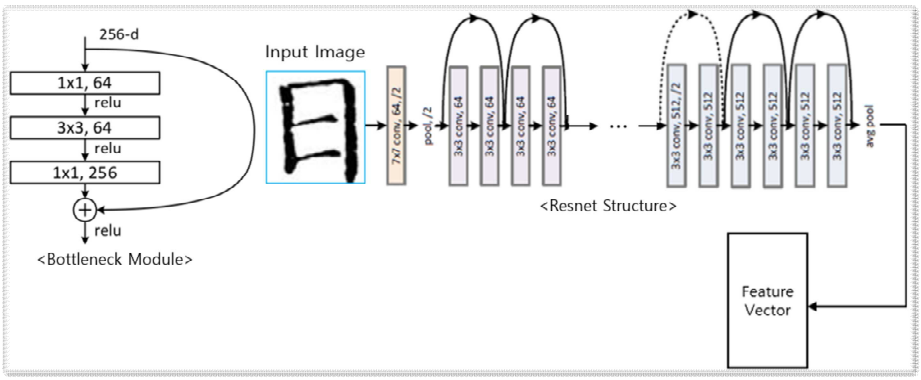
5.1 학습 모델

- 세그멘테이션 모델 (U-Net 기반)
  - 고문서 면단위 이미지가 U-Net 기반의 네트워크를 지나며 Heatmap으로 변환되고, Heatmap을 watershed 알고리즘을 통해 분리하여 낱자의 위치(바운딩박스)를 생성한다. 이때 Heatmap은 각 픽셀이 낱자의 중심에 위치할 확률을 표현



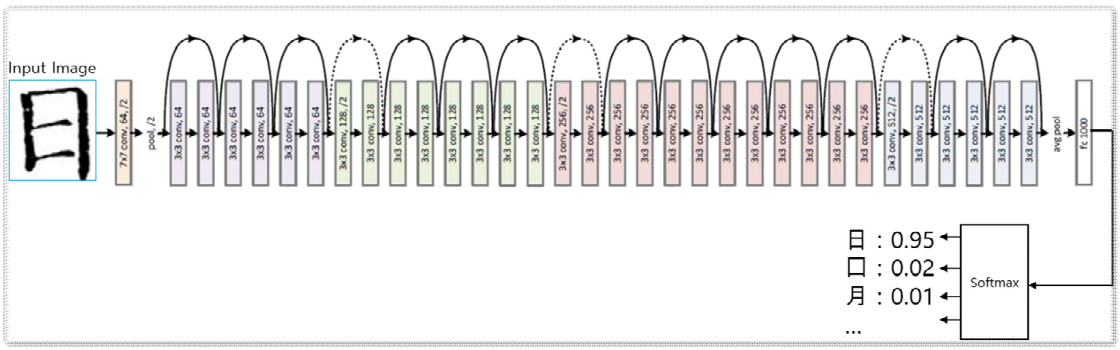
[그림 Ⅲ-11] U-Net 기반의 세그먼테이션 모델

- 클러스터 모델 (ResNet 기반)
  - 클러스터링 학습 모델은 Bottleneck을 적용한 ResNet을 기반으로 구성된다. 고문서 한자 낱자 데이터가 부족한 학습 데이터 구축 초기에는 한글 및 한자 오픈 데이터를 이용하여 모델을 학습한 후, 특징 추출부만 이용하여 한자 낱자 이미지의 특징 벡터를 추출하고 이를 각 글자의 특징벡터로 이용하여 특징 벡터간의 Cosine Similarity를 통해 글자간 유사도를 산출하고 유사도를 기준으로 유사 글자 클러스터를 구축



[그림 Ⅲ-12] ResNet 기반의 클러스터 모델

- OCR 인식 모델 (ResNet 기반)
  - OCR 인식 모델도 ResNet을 이용한다. 클러스터를 기반으로 작업자들이 구축한 OCR 용 한자 데이터를 이용해 학습하였으며 Softmax를 이용해 각 글자 이미지에 적합한 유니코드로 분류한다.



[그림 Ⅲ-13] ResNet 기반의 OCR 인식 학습 모델

5.2 서비스 활용 시나리오

- 구현된 활용 웹서비스에 접속한다. (로그인 불필요)
- 간단한 서비스 이용 안내문을 숙지한다.
- 한자가 포함된 이미지 파일을 업로드 한다.
- 업로드한 이미지 속의 한자가 인식되어 한자 텍스트로 변환된 결과를 웹 화면에서 확인한다.
- 필요할 경우 변환된 결과를 텍스트(*.txt) 파일로 다운로드 한다.

제3장 자유대화 AI 데이터

1 데이터 정보 요약

1.1 가이드 분류

대분류	오디오	중분류	자연어	소분류	WAV

1.2 데이터 정보

데이터 이름	• 자유대화 AI 데이터 - 자유대화 (일반남여) - 자유대화 (노인남여) - 자유대화 (소아남여, 유아 등 혼합) - 한국인 외래어 발화
데이터 요약	• 자유대화 (일반남여) - 녹음 인원 2,000명 이상, 4,000시간 음성 데이터 - 10대에서 50대 사이의 일반인 남녀의 발화 데이터 • 자유대화 (노인남여) - 녹음 인원 1,000명 이상, 3,000시간 음성 데이터 - 60세 이상의 남녀 발화 데이터 • 자유대화 (소아남여, 유아 등 혼합) - 녹음 인원 1,000명 이상, 3,000시간 음성 데이터 - 3세~6세, 7~10세 연령의 남녀 발화 데이터 • 한국인 외래어 발화 - 한국인이 발성한 외래어가 포함된 음성 데이터 - 녹음 인원 2,000명 이상 4,000시간 음성 데이터
데이터 출처	신규 제작

1.3 데이터 구축 개요

- 수집
  - 음성 데이터를 녹음할 녹음자를 모집 및 선별하고 작업 일정을 관리
  - 녹음자가 녹음 할 대화 시나리오(스크립트)를 작성
  - 온라인 / 오프라인 작업장을 구축하고 녹음자와 일정을 협의하여 녹취 진행하고 음성데이터 생성
  - 입력되는 음성 데이터 샘플링 주파수를 16kHz로 통일하여 설정하고, 44kHz 샘플링 rate로 녹음을 진행하는 경우 16kHz로 다운 샘플링 처리
- 가공
  - 데이터 가공을 위한 전사자 인력 모집 및 작업 환경 확보
  - 수집 과정에서 녹취한 음성데이터를 전사자가 정제 및 전사 수행하고 최종 전사 데이터 생성
- 검수
  - 데이터 검수를 위한 검수자를 인력 확보와 검수 규칙의 정비
  - 가공 과정에서 전사한 전사데이터를 검수자가 검수(1/2차에 걸쳐서 수행)를 하고 최종 학습 데이터를 생성
- 학습
  - 학습을 위한 학습모델을 정의하고 해당 알고리즘을 구현
  - 검수 과정에서 생성한 학습 데이터와 구현한 알고리즘을 바탕으로 학습 모델을 생성
  - 수집된 DB를 통해 BASE 엔진 기반으로 각 분야별, 음향, 언어모델 적응 학습을 수행하고, BASE 엔진 대비 인식 성능향상 여부의 유효성 검증을 진행
  - 각 분야별 응용 서비스에도 인식엔진을 제공하므로, 수요 업체의 서비스 시나리오의 유즈 케이스 문장을 통해 언어모델 적용 학습을 하여 언어모델 적응 학습 진행
- 응용 (응용 서비스 구성 시)
  - 학습데이터를 기반으로 한 응용 서비스를 정의
  - 정의된 서비스를 바탕으로 서비스를 개발하고 학습 모델이 생성되면 해당 서비스를 적용하여 응용서비스 개발
  - 문장을 통해 언어모델 적용 학습을 하여 언어모델 적응 학습 진행

1.4 구축 목적

- 인공지능(AI) 기반 한국어 음성인식 서비스 활성화를 위한 자유대화(일상대화) 지식 데이터 구축
- 실제로 사용하는 방대한 분량의 자유대화를 효과적으로 인식하기 위해 인공지능(AI) 기반 한국어 자유대화(일상대화) 데이터를 구축하며, 국민들에게 더욱 질 높은 인공지능(AI) 서비스를 제공 할 수 있는 양질의 학습데이터 확보하여 기술적 기반을 마련
- 영아/어린이, 노인층의 발화특성을 반영한 자유대화(일상대화)데이터를 구축하여, 음성인식 기반 인공지능(AI) 서비스 사용에 대하여 소외되는 계층이 없는 인프라 구성

※ 일반 성인남녀의 대화데이터를 사용할 경우 노인, 영아, 어린이 등의 발화의 특성을 고려되지 않기에 음성인식 기반 서비스가 정상적으로 제공되지 않는 가능성이 존재하며, 음성인식 기반 인공지능 서비스 사용에 대한 소외 계층 발생 가능

1.5 활용 분야

- 연구분야 : 음성인식, 음성언어처리, 자연어처리, 한국어 음성언어연구, 신호처리 등
- 산업분야 : 온/오프라인 기반의 음성인식, AI비서, Voice BOT, Voice Command & Control, AI 로봇, 음성인식기반 키오스크

1.6 유의 사항

- 데이터 구축 시 외부 데이터 활용 하는 경우 법적/사회적 이슈가 없도록 데이터 무상 확대제공에 대한 라이선스를 득하여 확약 처리
- 데이터 제공 업체 또는 기관의 대표 명의 공문 혹은 확약서로 명문화
- 데이터 지원 확약서 작성 시 양식에 따라 데이터명, 데이터 제공 기간, 데이터 활용 과제명 등의 정보를 명시

2 데이터 획득 및 정제

2.1 원시 데이터 선정

- 원시데이터 수집을 위한 대화시나리오 선정
  - 작성 필요성: 발화 수집목적에 부합하는 대화주제 및 원고개발을 통해 음성수집 목표 달성
  - 작성절차: 수집기준 선정 → 수집 → 정제 → 분류 → 발화원고 및 주제작성

작성절차	일반남여	노인남여	소아남여, 유아 등 혼합	한국인 외래어 발화
수집기준 선정	수집항목 선정 · 일상 대화 주제 분류 · 대화 상황 및 장소, 공간 선정	수집항목 선정 · 일상 대화 주제 분류 · 대화 상황 및 장소, 공간 선정	수집항목 선정 · 3-5세 누리과정 생활주제, MCDI-K 기반 선정 · 국립국어원〈초등학생을 위한 표준 한국어 의사소통〉참고	수집항목 선정 · 외래어 단어 및 해당 단어를 사용한 용례 발화 수집 · 콘텐츠와 관련한 인명, 지명, 제목 포함 수집
수집	생활 어휘 수집 · 인터넷, 유튜브, SNS 검색 · 웹 크롤링을 통한 수집 · 한국어 교재분석	노인 어휘 수집 · 장년층 대상 방송 프로그램 대본 분석 · 웹 크롤링을 통한 수집 · 방언 관련 논문 분석	소아 어휘 수집 · EBS 어린이 방송 대본 분석 · 소아대상 유튜브 검색 · 초등학교 국어교과서, 동화책 주제, 소재, 성취기준 분석	외래어 어휘 수집 · 인터넷, 유튜브, SNS, 블로그를 검색 · 웹 크롤링을 통한 수집
정제	생활어휘 정제 · 생활어휘 사용빈도 분석 · 민감한 이슈 발언 제외	노인 어휘 정제 · 노인 어휘 사용빈도 분석 · 대상연령 적합 어휘 선별	소아 어휘 정제 · 소아 어휘 사용빈도 분석 · 대상연령 적합 어휘 선별	외래어 정제 · 외래어 사용빈도 분석 · 인명,지명 최소화
분류	생활어휘 분류 · 대화주제별 카테고리 분류 · 비문법적 표현, 신조어, 말줄임, 반복 등 대화 선정	노인 어휘 분류 · 대화상황별 분류 · 비문법적인 표현, 말 줄임, 노인 특유의 부정확한 발음 어휘 선정	소아 어휘 분류 · 대화상황별 분류 · 비문법적인 표현, 말 줄임, 소아 특유의 부정확한 발음 어휘 선정	외래어 분류 · 유래어 분류 · 주제별 카테고리
발화원고 주제 작성	대화 시나리오 작성 · 성별, 연령별 주제 선정 · Small Talk대화, Voice 챗봇 등 상황별, 장소 별 작성	대화 시나리오 작성 · 억양 및 단어 사투리 반영 · Small Talk대화, Voice 챗봇 등 상황별, 장소 별 작성	대화 시나리오 작성 · 3~6세, 7~10세 주제 선정 · 낭독 발화, 주제별대화, 자유발화, Small Talk대화 등 원고 작성	대화 시나리오 작성 · 유래어 분류 · 주제별 카테고리 · 성별, 연령별 어휘

[그림 Ⅲ-14] 원시데이터 선정 절차 및 수집항목

- 일반남녀 발화목록 작성방안
  - 일반 남여 언어 특징을 고려하여 대화 주제, 상황 별 대화 시나리오 작성
  - 대화 시나리오를 보고 읽는 낭독 발화에 대한 대화 시나리오 작성
  - 자유발화 : 제시된 주제를 바탕으로 다수의 사용자가 자유롭게 채팅을 하는 형태로 대화를 진행하도록 자연스럽게 대화를 할 수 있도록 지인 또는 가족간 대화 방을 구성할 수 있도록 대화방 내 초대 기능 활용

- 발화분량 : 총 2시간의 발화분량으로 진행하나, 발화자의 노인 발화자의 건강 및 집중력을 고려하여 녹음 시간 단위를 1시간 단위로 제한하거나 한 번에 수집할 수 있는 분량을 제한

● 노인남녀 발화목록 작성 방안

- 노인 연령대의 언어 특징을 고려하여 대화 주제, 상황 별 대화 시나리오 작성 필요
- 대화 시나리오를 보고 읽는 낭독 발화에 대한 대화 시나리오 작성
- 발화분량 : 총 2시간의 발화분량으로 진행하나, 발화자의 노인 발화자의 건강 및 집중력을 고려하여 녹음 시간 단위를 1시간 단위로 제한하거나 한 번에 수집할 수 있는 분량을 제한하는 것을 고려

〈표 Ⅲ-6〉 발화목록 작성 방안(일반/노인 남녀)

발화종류	발화분량	발화방법	발화 내용
낭독 발화	1시간 1200단어/ 800문장	다양한 일상 대화를 2사람 이상이 대화 화는 대화체 주제 별로 나누어 역할 별로 읽게 하거나, 단문의 내용을 읽게 함	〈문장 낭독〉 - 한국어의 음운이 고루 실현되도록 작성한 문장 - 각 대화 스크립트 별 PBS(Phonetically Balanced Sentences) 문장을 추가 〈문단 낭독〉 - 정보문이나 스토리 등의 내용으로 작성한 문장
자유 발화	1시간	주제만 제시하고 자연스럽게 자발적으로 대화를 자유롭게 끌어가도록 함	〈자유 발화〉 - 제시된 주제에 대해서 자유롭게 대화 발화 ex) 가족 소개, 아끼는 물건, 애완동물 등

● 소아남녀 발화목록 작성 방안

- 연령대의 인지 발달, 언어 수준을 고려하여 대화 주제, 상황 별 대화 시나리오 작성 필요
- 대화 시나리오를 보고 읽는 낭독 발화에 대한 대화 시나리오 작성
- 소아의 경우 문장의 비율을 50%이상으로 녹음 지문을 확보
- 발화분량 : 총 2시간의 발화분량으로 진행하나, 발화자의 집중력과 녹음 상태를 고려하여 녹음시간 단위를 1시간 단위로 제한하거나 한 번에 수집할 수 있는 분량을 제한하는 것을 고려

〈표 Ⅲ-7〉 발화목록 작성 방안(소아남녀)

대상	대화 주제	발화목록 예시
3~6세	• 교육부 3~5세 누리과정의 생활주제를 중심으로 대화 주제 및 내용 선정 • 한국의 영유아가 사용하는 어휘를 설계한 MCDI-K (MacArthur Communicative Development Inventory-Korean)를 참고하여 단어를 1차적으로 선정하고 대화 주제 반영	〈낭독 발화〉 • 단어(그림 보고 말하기) • 문장(듣고 따라하기, 질문과 대답) • 주제별 발화 ex) 좋아하는 것 말하기, 색깔 말하기 등 • 영상을 보고 말해 보기, 챗봇과 대화하기 등
7~10세	• 국립국어원 기획 [초등학생을 위한 표준한국어 의사소통] 교재를 중심으로 대화 주제 선정 • 교육과정 국어 초등학교 1-2학년군 및 3-4학년군 성취기준 및 국어교과서 참고	〈낭독 발화〉 • 문장(초등 교과서나 동화책) • 문단(초등 교과서나 동화책) • 주제별 발화 ex) 자기소개, 물건 사기, 여행 등

● 한국인 외래어 발화목록 작성방안

- 한글이나 국어로 대체되어 쓰이는 경우가 더 많은 어휘 제외 (인위적으로 정한 순화어는 예외)
- 비속어, 은어 등 제외
- 아직 일반화되지 않은 외국어의 한글 발음이나 표기 제외
- 발화분량 : 총 2시간의 발화분량으로 진행하나, 발화자의 집중력과 녹음 상태를 고려하여 녹음 시간 단위를 30분 단위로 제한하거나 한 번에 수집할 수 있는 분량을 제한하는 것을 고려
- 외래어 단어 및 해당 단어를 사용한 문장단위의 용례발화 포함하여 수집 (모든 문장이 영어로된 문장은 20% 비율 이하로 제한)
- 콘텐츠(영화, 노래)와 관련된 인명, 지명, 제목 포함

〈표 Ⅲ-8〉 발화목록 작성 방안(외래어)

발화 종류	세부 내용
외래어 어휘 수집	• 인터넷, SNS, 블로그를 검색하여 최대한 많은 외래어 수집 • 웹 크롤링/클리닝을 통한 수집 (Python, Perl 등의 언어로 택스트 프로세싱 대화 시나리오를 생성하여 진행)
외래어 목록 정제	• 외래어 사용빈도 분석: 가능한 한 대용량의 그리고 복수의 말뭉치(코퍼스)를 이용하여 각 외래어 마다 사용빈도를 분석 • 사용빈도를 정렬(sorting)하여 일정 값(threshold)을 기준으로 제한 • 컷오프 기준은 데이터의 양, 수집 시간 등을 고려

발화 종류	세부 내용
외래어 유형 태깅 (categorization and tagging)	• 유형 설정: 인명, 지명, 상품명, 프로그램 명, 전문 용어, 스포츠 용어 등 • 유래어 확인: 그 외래어의 유래 언어 검색: 영어, 불어, 일본어 등 • 유형과 유래어에 관한 태깅: 준자동 태깅을 진행하거나, 수작업으로 진행

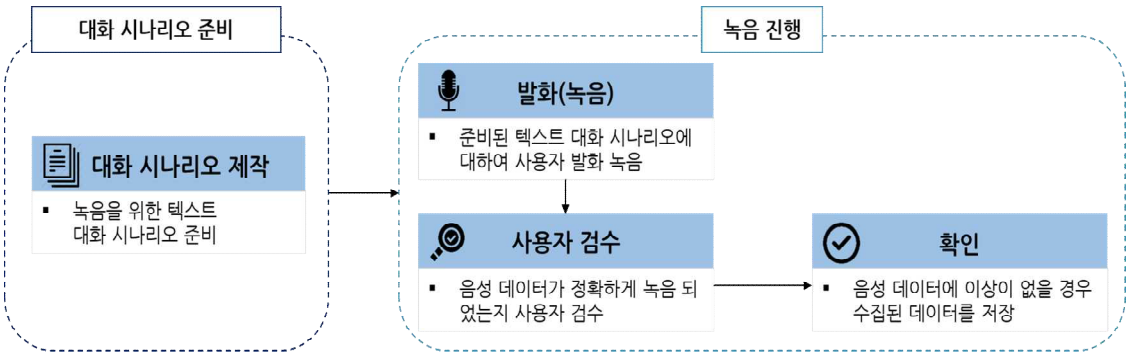
- 대화 시나리오 작성 시 고려사항
  - 시나리오 작성 시 음성 인식에 적합하도록 Phone balance를 고려하여 작성
  - 대화 시나리오에 따라 음성 데이터의 정합성 및 다양성에 영향을 끼치므로, 과제 시작 후 대화 시나리오 작성 내용 검토
  - 일정 길이 이상의 문장으로 이루어져 있는지, 단순 단어 형식인지 여부 검토 (발화의 길이가 밸런싱 있게 녹음이 될 수 있도록 고려하여 대화 시나리오를 구성)
  - 시나리오 작성 건이 다른 콘텐츠(소설, 영화 및 드라마 대본)의 내용을 그대로 차용했는지 여부 등을 검토하여 저작권 위반 여부 별도 검증하고, 시나리오 작성 시 타 기관의 콘텐츠가 필요한 경우 협의를 통해 데이터 제공 확약서를 작성하고, 해당 콘텐츠 제공받아 작성
  - 시나리오 작성 과정에서 실제 녹음작업을 수행할 업체의 실무자를 통해 음성인식의 적합한 형태인지를 검증
  - 동일한 내용을 녹음하여 중복되는 데이터가 다수 발생 않도록 충분한 시나리오를 확보하며 시나리오가 부족하지 않도록 녹음이 진행되는 과정에서도 추가적으로 시나리오를 작성될 수 있도록 확인

2.2 규제관련 사항

- 관련 내용 없음

※ 본 데이터 가이드라인 상에는 규제 관련 내용이 없지만, 데이터 구축 시 규제 관련 내용 확인 및 검토 필요

2.3 획득 및 정제 절차



[그림 Ⅲ-15] 발화 데이터 획득 및 정제 절차

- 발화 : 녹음하기 버튼을 통해 준비된 대화 시나리오를 사용자가 발화
- 사용자 검수 : 녹음된 음성을 들어보고 수정이 필요한 경우 다시 녹음
- 확인 : 녹음된 음성데이터가 이상이 없는 경우 저장

2.4 획득 및 정제 기준

- 녹음 수집 데이터 적/부 판별 기준
  - 샘플링 주파수 16kHz
  - 16Bit Resolution 기준으로 오차 ± 30,000 이상의 샘플이 20% 이상인 데이터는 불량으로 판정
  - 발화와 발화 사이의 무음 구간이 2초 이상인 데이터는 불량으로 판정
  - 발화문장 앞뒤의 불필요한 무음은 제외 (단 발화 앞뒤로 100~200msec 정도의 묵음은 음성 인식을 위해 필요)
  - 데이터 검수 과정에서 수집데이터의 비율 중 10% 정도의 데이터 불량이 발생할 것으로 예측 (1차 검증 : 전수검사 / 2차검증 : 선별검사)
  - 데이터 가공 및 정제 과정에서 음원의 왜곡이 없도록 진행
  - 불량 데이터가 발생할 것과 추가 수요가 발생할 것을 감안하여 고려하여 원시데이터(녹취데이터) 수집 과정에서 구축데이터 목표 분량의 5~10%를 추가 수집
- (클리핑 방지 가이드)

- 녹음실에서 음성 DB 수집 시, PC 녹음 매뉴얼의 녹음 가이드 정책을 따라 작업 수행
- 대화 앞뒤의 Silence Margin 없이 발성을 방지하기 위해 음성 녹음 시작 지점에 0.5초 구간의 Red Zone을 두어, 이 후에 발성하도록 시각적으로 표시
- Waveform 형식으로 시각적으로 클리핑이 될 정도로 모든 음성을 크게 발성하는 경우는 재녹음을 진행
- 근래의 Android, iOS 단말의 코덱 성능으로는 32,000 이상으로 크게 녹음되는 경우, 찢어지는 소리 없이 AGC를 이용하여 음성인식으로 입력으로 사용 가능
- 클리핑된 음성이 녹음되는 경우, 음성인식 학습 측면을 고려하여 다양한 조건의 음성 DB 확보가 강인한 인식 성능 보장 가능

3 어노테이션/라벨링

3.1 어노테이션 / 라벨링 절차

● 음성인식(전사) 절차

① 입력되는 음성 데이터의 샘플링 주파수 설정: 16kHz

② 녹음 데이터의 지정

- 전사를 위한 음성 파일과 발화 텍스트 파일을 지정

③ 전사자는 수정 edit 창에 음성을 듣고, 녹음자의 발성이 정확하게 철자전사 되어 있는지 확인 후, 틀린 발성이 있는 경우는 텍스트를 수정

④ 음성인식 학습을 위해 사용되는 데이터이기 때문에, 음성 구간의 앞과 뒤에 1초정도의 무음 구간이 존재해야 하나, 지정된 시간보다 짧은 무음 구간이 있을 경우에는 데이터의 무음 구간을 복사하여 붙여 넣을 수 있는 기능 제공

⑤ Studio에서 녹음DB 이외의 실 환경에서 수집 받은 음성데이터에는 외부 잡음 등이 포함되어 있으니, 외부 잡음 Filler Noise를 철자 전사 정보에 추가

- (NO:) 내일 오후 세시에 예약해 주세요

⑥ 자유 발화인 경우, 녹음자도 모르는 사이에 발성하는 어', '음', '글세' 등도 음성인식 학습에 필요한 정보이기 때문에 간투어 Filler Pause 정보를 철자 전사 정보에 추가

- (FP:음) 내일 오후 세시에 예약해 주세요

⑦ 유아, 노인들이 발성할 경우, 전사자가 들었을 때 명확하게 철자전사 작업을 하지 못할 음성 전사는 발성오류 정보를 추가

- 내일 오후 (SP:세시에) 예약해 주세요 : 세시, 네시 가 명확하게 들리지 않는 경우

⑧ 마이크로폰 근접에서 발성하는 경우, 녹음자의 들숨, 날숨, 웃음 소리 등이 녹음되는 음원의 경우는 화자 잡음 Speaker Noise를 철자 전사 정보에 추가

- (SN:) 내일 오후 세시에 예약해 주세요 (SN:)

⑨ 모든 입력 버튼, Play command는 마우스 움직임을 통한 click 실행보다는 시간 단축을 위해 단축키를 설정하여 전사 작업을 수행

- Play : CTRL + Space
- 화자잡음 : SHIFT + F1

⑩ 일반적으로 정의된 전사 규칙 외에 정보 입력이 필요할 경우, MEMO 창에 입력

⑪ 파일별로 전사 작업이 완료되면, 음성데이터 파일명과 동일한 TRS 파일을 생성

```
[Text Information]
The Original EPD Start=0
The Original EPD End=440
The Original Text= 내일 오후 세시에 예약해 주세요
The Modified EPD Start=0
The Modified EPD End=440
The Modified Text= (FP:음) 내일 오후 세시에 예약해 주세요
Memo=
```

[그림 Ⅲ-16] 음성인식-전사 정보 예시

⑫ 전사자와 검수자는 동일한 전사툴을 사용하며, 검수자가 검수할 경우에는 TRS 파일을 열어 수정된 철자 정보를 열고, 수정 작업을 진행

⑬ 스튜디오 녹음 수집 외 온라인 등 다른 방법으로 수집되는 정보를 각 참여기관과 협의하여 진행

3.2 어노테이션 / 라벨링 기준

- 개요
  - 숫자, 영어, 기호를 사용하지 않고 한글로만 전사
  - 전사 시 전사규칙과 관련된 기호 이외는 비사용
  - 입력 가능한 기호 : (SP:), (FP:), (SN:), (NO:)
  - 표준발성에서 벗어나거나 같은 전사에 대하여 두 가지 이상 발음이 가능한 경우 발음전사 표기
  - 철자전사 : 표준어법에 맞게 표기하고, 음성인식의 언어모델링 등을 주된 목적으로 함
  - 발음전사 : 발성된 내용을 소리 값에 최대한 가깝게 표기하고, 음성인식의 음향모델링을 주된 목적으로 함
  - 단어의 앞과 뒤에 거의 붙어 발생된 잡음은 단어와 분리하여 표기
  - 잡음이 있는 상황에서 사람에게서 발생하는 잡음(입술소리, 숨소리)은 명확히 구분될 정도로 큰 것만 표기
  - 화자 잡음, (SN:) : 웃음소리, 숨소리, 입술소리
  - 외부 잡음, (NO:) : 녹음자 이외의 주변 잡음, 음악소리
  - 띄어쓰기는 한글 맞춤법에 맞도록 하되, 표준어법으로 명확히 결정할 수 없는 경우에 띄움
- 숫자표현
  - 기본적으로 숫자는 모두 숫자 기호가 아닌 한글로 표현
  - 한국어의 경우 십진 단위로 띄어쓰기
  - 숫자를 하나씩 발음한 경우에 띄어쓰기
  - 단위를 나타내는 '년', '월', '일', '시', '분' 등은 붙여쓰기

    예) 오대 그룹이 모여/자동차 다섯 대를
    이십 사시간(24시간)/스물 네시간(24시간)
    팔 육 공에 이 사 삼 칠(860-2437)
    십 사시(14시)/열 네시(14시)
    천 구백 구십 구년에 (1999년에)
  - 숫자만으로 이루어진 기념일 등 특정 의미가 있는 단어들을 숫자 단위로 띄어쓰기

    예) 팔 일 오 (8.15) , 사 일 구 (4.19)
    오 칠 오 공 부대(5750부대)

- 간투어표현
  - 발성자가 다음 발성을 준비하기 위해서 소요되는 시간을 벌기 위해서 발성하는 것으로 의미 없음
  - 간투어를 포함하여 (FP:**) 로 표기

    예) (FP:에)/(FP:아)/(FP:그)/(FP:어)/(FP:음)/(FP:저)/(FP:저기)/(FP:으) /(FP:응)
- 약어/외래어 표현
  - 약어의 형태의 알파벳을 발화하는 경우, 붙여쓰기

    (하단 알파벳 한글 표기 참조)

    예) 케이비에스 (KBS)/에이티엔티 (AT&T)
  - 된소리 나는 외래어는 표준어 로 철자전사로 표기

    예) 스타벅스(스타벅쓰), 서비스(써비쓰), 센터(쎈터)
  - 우리말로 표기하여 자연스러운 것은 통상적인 한글 표현으로 표기

    예) 뉴욕, 시카고, 파티
  - 버스, 핸드폰, 모바일, 인터넷, 호텔

알파벳	한글 표기	알파벳	한글 표기
A	에이	N	엔
B	비	O	오
C	씨	P	피
D	디	Q	큐
E	이	R	알
F	에프	S	에스
G	지	T	티
H	에이치	U	유
I	아이	V	브이 / 비
J	제이	W	더블유
K	케이	X	엑스
L	엘	Y	와이
M	엠	Z	지/제트

〈표 Ⅲ-9〉 알파벳 한글 표기

- 비표준발음 표현

  - 어미 비표준 발음으로 들리는 경우 표준어 형태의 철자전사로 표기

    예) 같아요 ← 같아여, 같애요
    했고 ← 했구
    했고요 ← 했구여, 했고여, 했구요
    입니다 ← 입니더
    밥을 먹었고요 ← 밥을 먹었구요, 밥을 먹었구여, 밥을 먹었고여

- 기타

  - 축약 발성은 표준어 형태의 철자전사로 표기

    예) 안녕하세요 정선희입니다 ← 안녕하세요 정선흽니다

  - 알아 듣기 힘든 발음, 발성과 동시에 발성된 잡음 처리

  - 화자가 발음한 내용을 잘 알아 듣기 힘들 때, (SP:**)로 표기

    예) 나는 (FP:이럴끄) 그것을 해결하였다

  - 발성과 동시에 발생하는 외부 잡음은 (NO:**)로 표기

    예) 기차 타는 (NO:곳이) 어디입니까 (‘곳이’발성할 때 외부 잡음이 크게 섞임)

  - 반복 발성이나 잘못된 발성은 (SP:**)로 표기

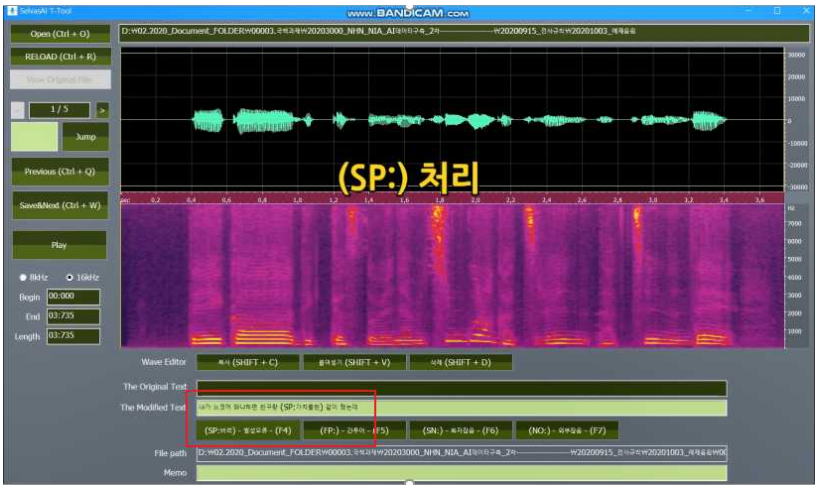
    예) 아침에 (SP:학교) 학교에 갔다

  - 방언에 해당하는 발성은 발음 전사로 표기

    예) 핵교 (학교의 방언)

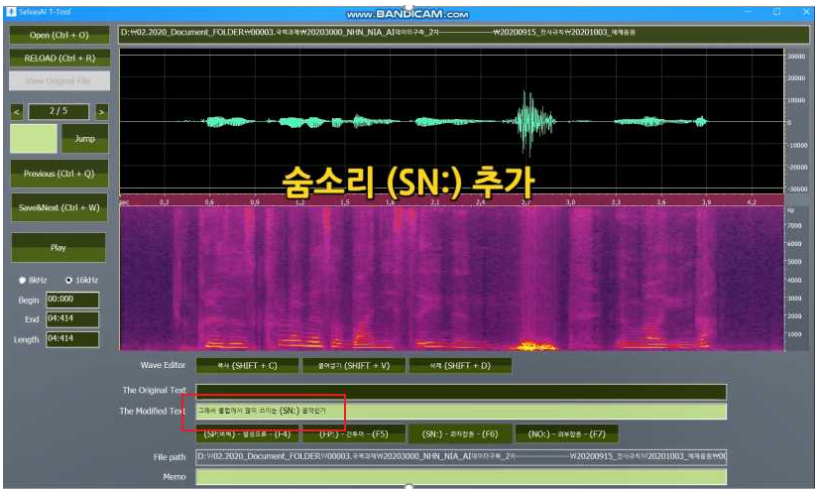
  - 대화체 문장은 문장 자체가 이상하더라도 발음 전사

  - 버벅 거림(SP) 표기



[그림 Ⅲ-17] SelvasAI T-Tool 사용예시 - 버벅 거림 (SP:) 표기

- 화자 잡음(SN) 표기



[그림 Ⅲ-18] SelvasAI T-Tool 사용예시 - 화자 잡음 (SN:) 표기

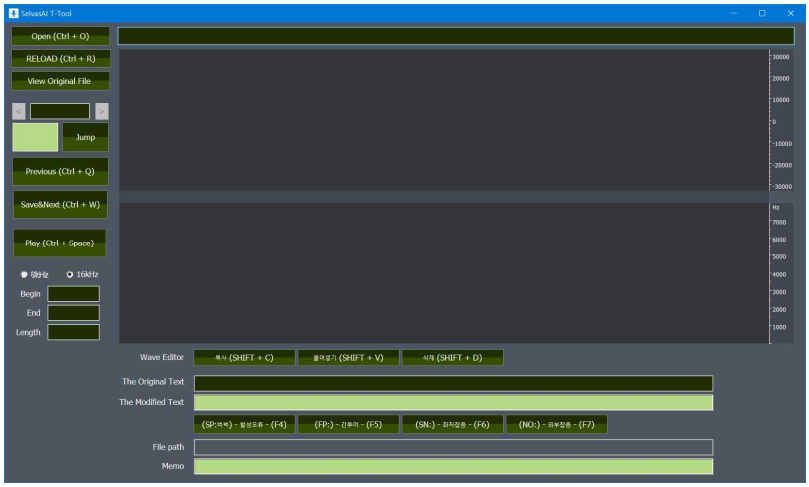
3.3 어노테이션 / 라벨링 교육

- 관련 내용 없음

3.4 어노테이션 / 라벨링 도구 및 사용법

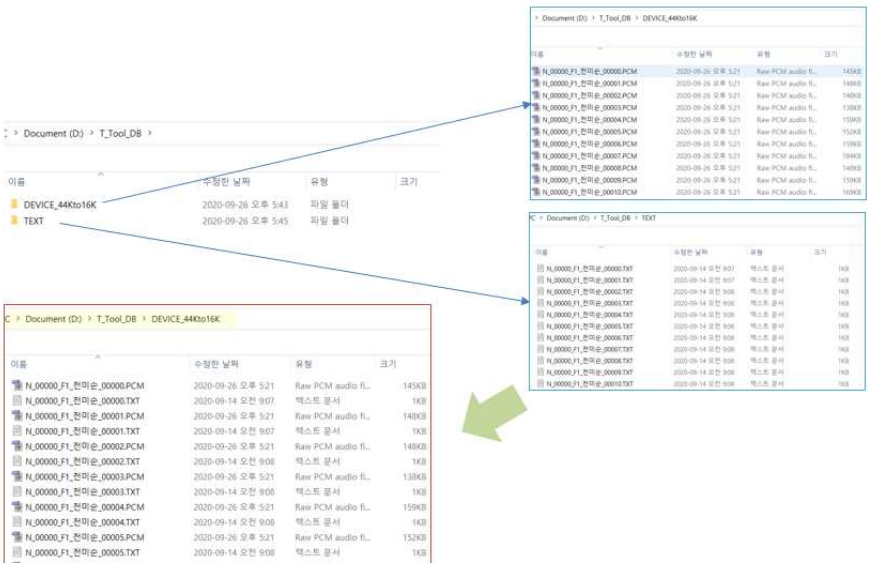
1) 어노테이션/라벨링 도구 1 : 셀바스 AI

- 개요
  - 녹음실에서 지정된 발화 script를 보고 발성
  - 데이터 정제 작업 : SelvasAI T-Tool을 이용하여 녹음자가 발화한 음원의 script를 확인, 수정하는 작업
  - silence 음성 구간 삽입, 삭제, Script 수정, 외부잡음, 화자잡음, 간투어 정보 추가문자열로 대치함



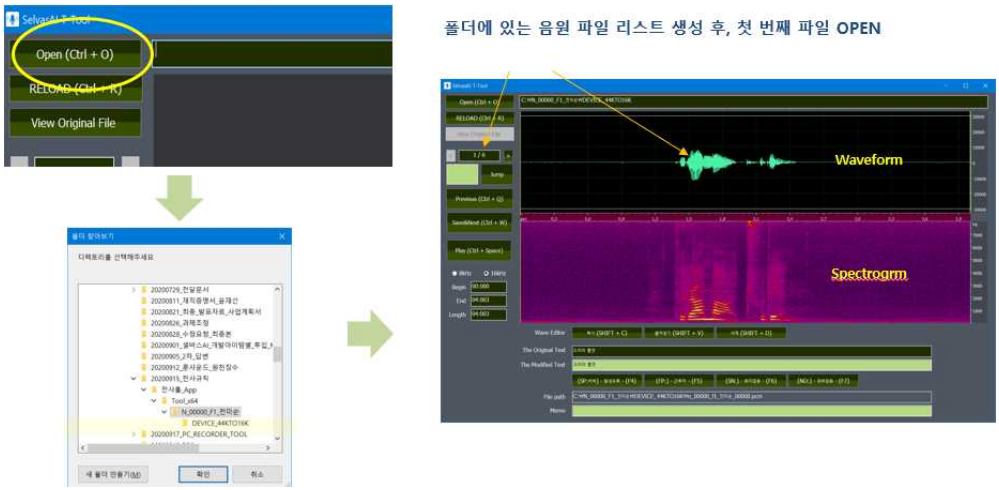
[그림 Ⅲ-19] SelvasAI T-Tool 화면

- 사전작업
  - DEVICE44Kto16K : *.PCM 음원 폴더
  - TEXT : 대본 script
  - 대본 script text를 DEVICE44Kto16K 폴더에 복사



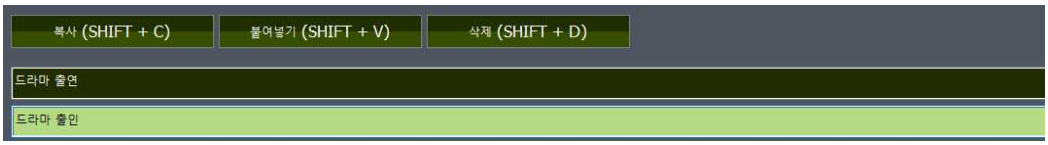
[그림 Ⅲ-20] 사전작업

- 전사도구(SelvasAI T-Tool) 사용법
  - OPEN 버튼을 click 하여 전사할 음원, scrip가 있는 폴더를 선택



[그림 Ⅲ-21] 음원, script 폴더 선택

- 전사작업
  - The Modified Text 영역에 script 수정
  - 드라마 출인 이라고 발성 ? “드라마 출인“ 으로 script 변경

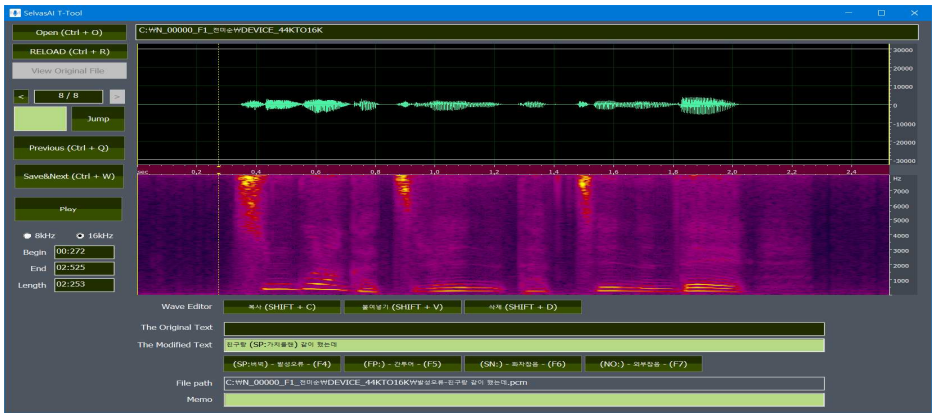


[그림 Ⅲ-22] 전사 작업

- 작업 내용 저장 및 다음 음성 파일 OPEN
  - Save&Next button click or CTRL+W

[그림 Ⅲ-23] 저장 및 다음 음성 파일 OPEN

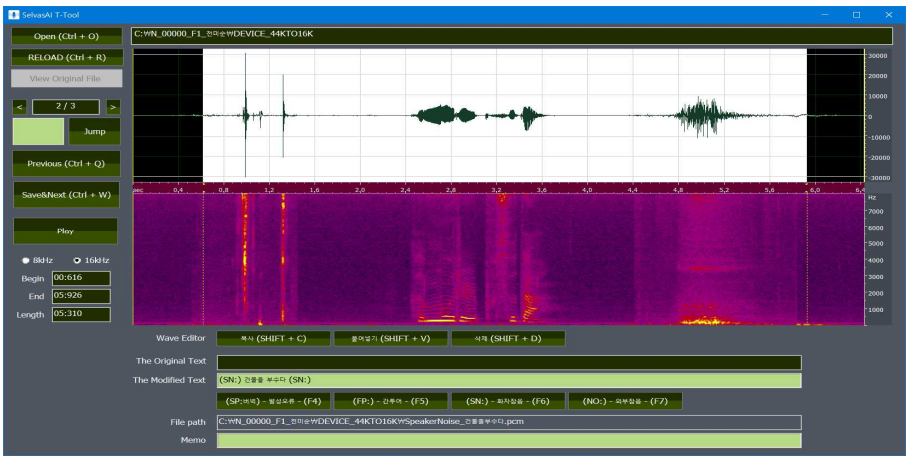
- 잡음, Filter 정보 추가 방법
  - 발성 오류, 버벅 거리는 음원은 소리나는대로 전사하고 SP Filler 추가



[그림 Ⅲ-24] 잡음 및 Filter정보 추가

- 입술 소리, 숨소리, 웃음 소리 는 Speaker Noise (SN:) 추가
  - 입술 소리나 숨소리는 대부분 녹음실 환경에서 근접 마이크로 녹음시에만 녹음됨

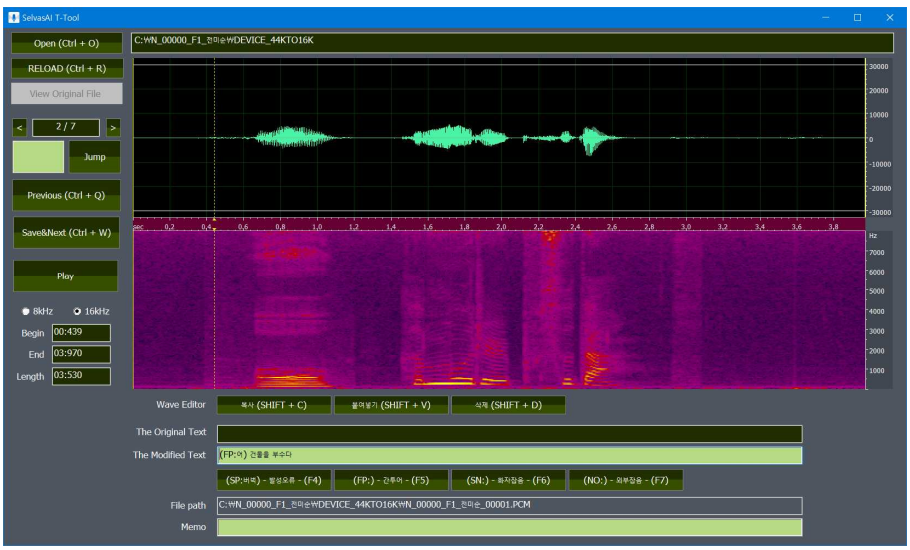
    예) (SN:) 건물을 부수다 (SN:)



[그림 Ⅲ-25] 잡음 및 Filter정보 추가 - Speaker Noise (SN:)

- 문장 시작 전에 습관적으로 발성하는 뭐, 음, 어 같은 발화가 녹음된 경우 Filler Noise (FP:음), (FP:어) 추가

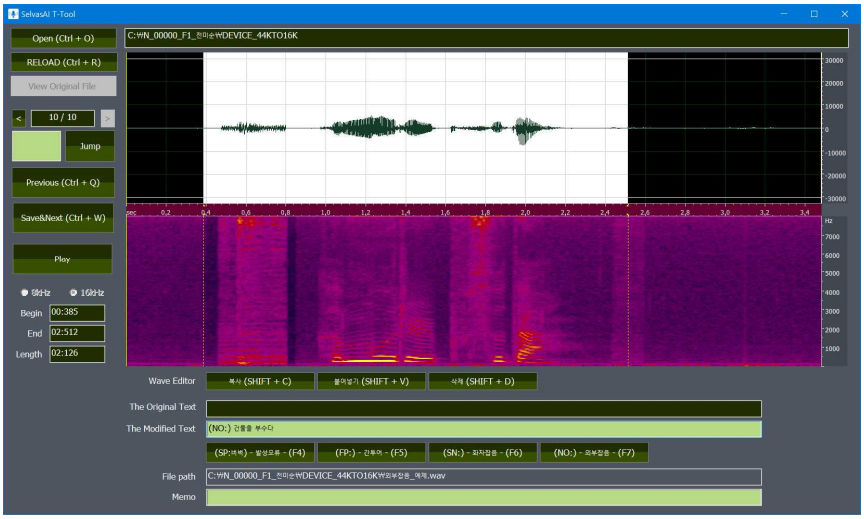
  예) (FP:어) 건물을 부수다



[그림 Ⅲ-26] 잡음 및 Filter정보 추가 - Filler Noise (FP:어)

- 발화자 이외의 녹음 중외부 잡음이 들어 온 경우, (NO:) 추가

  예) (NO:) 건물을 부수다



[그림 Ⅲ-27] 잡음 및 Filter정보 추가 – 외부 잡음(NO:)

- 파일별 전사 결과 저장

  - 음원 저장 폴더와 동일한 이름_OK 폴더 생성

[그림 Ⅲ-28] 그림 저장 폴더(_OK) 생성

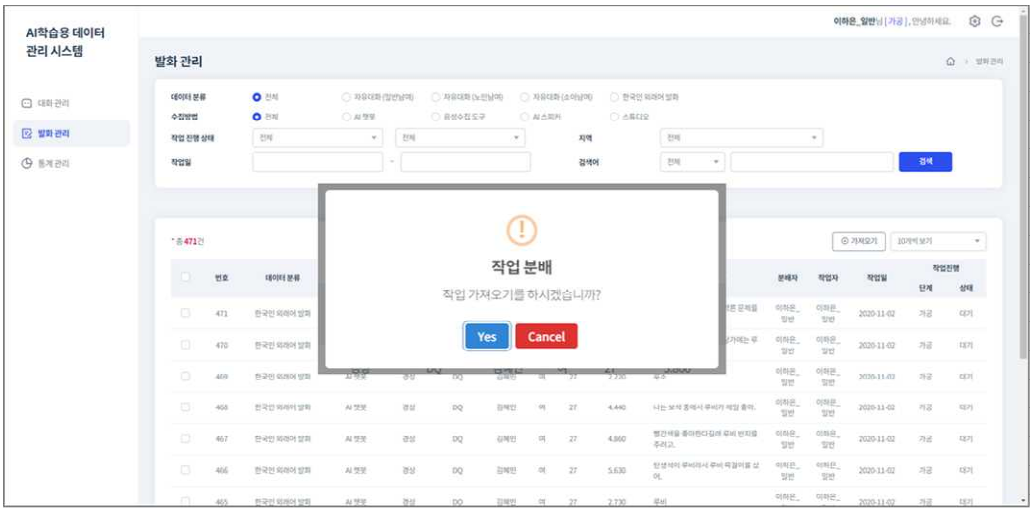
  - **_OK 폴더 아래에 음원과 전사 정보 파일(.TRS) 저장

```
[Text Information]
The Original EPD Start=0
The Original EPD End=348
The Original Text=건물을 부수다
The Original Sample Rate=16000
The Modified EPD Start=0
The Modified EPD End=348
The Modified Text=(NO:) 건물을 부수다
Memo=
Voice Type=
```

[그림 Ⅲ-29] 전사 정보 파일(.TRS) 예시

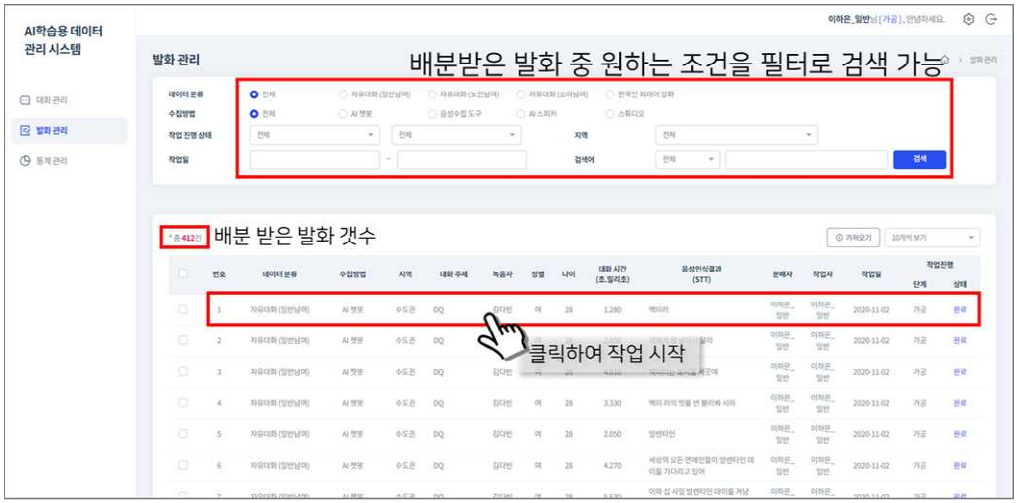
2) 어노테이션/라벨링 도구 2 : NHN다이퀘스트

- 개요
  - 데이터 가져오기 및 작업 분배 기능을 통한 검수 데이터 배분
  - 다수의 검수자가 동시에 병렬 검수 작업 가능
  - 검색 및 필터링 기능을 통한 검수 대상 발화 식별
  - 음성데이터 재생 및 가공(전사) 문서 비교하여 수정
  - 반려 기준에 따라 검토하고, 기준을 충족하지 못할 경우 반려 처리
- 발화 가져오기
  - 가져오기시 10분(약 150 문장)분량 데이터 배분
  - 최대 10분 분량만 배분 가능하며 작업 완료 전 추가로 가져오기 누를 시 작업 진행된 분량 만큼만 추가



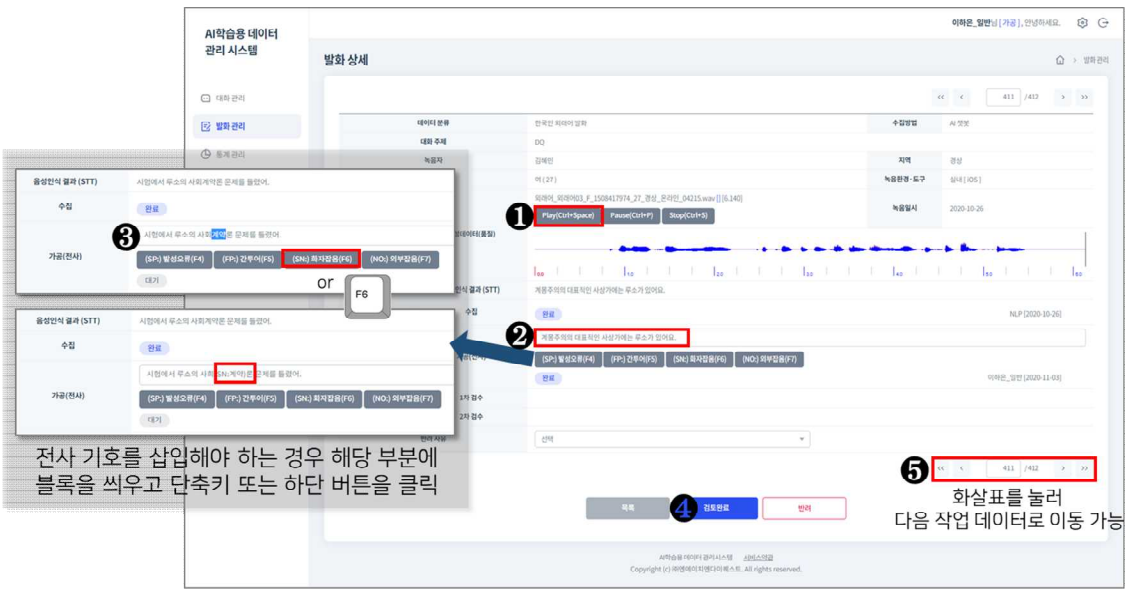
[그림 Ⅲ-30] NHN다이퀘스트의 발화 가져오기 화면

- 작업 시작
  - 배분 받은 발화 중 원하는 조건을 필터로 검색 가능
  - 배분 받은 발화 개수 확인 및 발화 선택하여 가공 작업
  - 가공 대기 상태로 검색 시 작업 대기 중인 데이터만 검색
  - 작업진행 단계는 ‘분배 후 작업전: 가공 대기’, ‘검토 완료: 가공 완료’로 구분



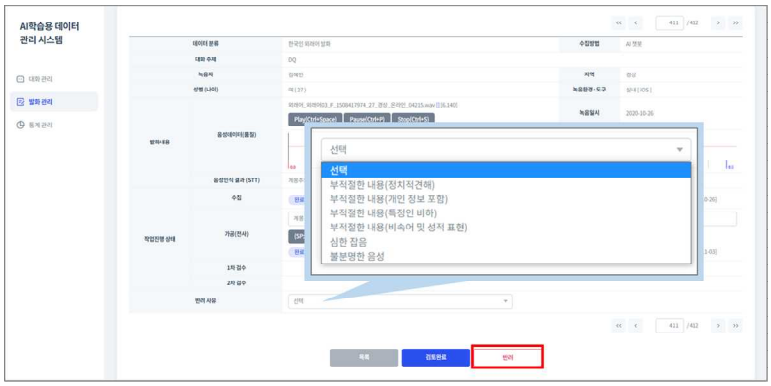
[그림 Ⅲ-31] NHN다이퀘스트의 작업시작 전 화면

- 가공(전사)
  - Play 버튼(Ctrl+Space)을 눌러 녹음된 음성데이터를 듣고 가공(전사) 문장과 비교하여 수정
  - 전사 기호를 삽입해야 하는 경우 해당 부분에 블록을 씌우고 단축키 또는 하단 버튼을 클릭하여 처리
  - 검토완료 버튼 클릭하여 저장
  - 화살표를 눌러 다음 데이터로 이동



[그림 Ⅲ-32] NHN다이퀘스트의 발화 가져오기 화면

- 반려
  - 반려 항목의 경우 반려 사유를 선택하고 검토 완료 버튼이 아닌 반려 버튼 클릭
  - 반려 기준은 '정치적 견해 포함', '개인 정보 포함', ' 특정인 비방', '비속어 및 성적 표현', '심한 잡음', '불분명한 발음' 등이 있음

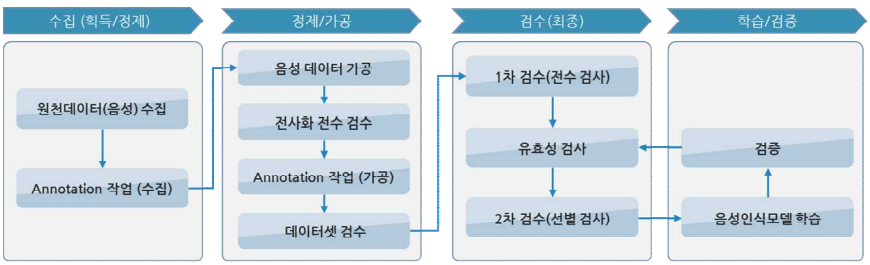


[그림 Ⅲ-33] NHN다이퀘스트의 발화관리 반려 화면

4 데이터 검수

4.1 검수 절차

- 검수프로세스



세부	수집	정제/가공	검수(최종)	학습/검증
검수 단계	• PCM 형태의 음성파일 수집 • Annotation 작업(수집)을 통해 노이즈 정도 체크 • 손상된 음성, 인식하기 어려운 음성 데이터 오류 체크 및 제외 • 분류기준(카테고리)별 음성 분류 유무 체크 • 수집 데이터 작업자 자체 및 Cross Check	• 음성 데이터 텍스트화 작업 • 전사화 전수 검수를 통해 음성 파일 및 전사 텍스트 정합성 체크 및 민감 이슈(정치, 개인정보 등) 발언 저작 도구 Tool 활용하여 제외 처리 • 전사 데이터 라벨링 작업을 통해 분류 기준 별 데이터 매칭 여부 체크 • 전사 데이터에 대한 자체 및 크로스 검증을 통해 어노테이션 검증 결과 일치율 96% 이상이면 과반수에 의해 True 처리	• 가공 완료된 최종 산출물(학습데이터) 전수 검수 • 라벨링 누락 및 정합성 체크 • 전체 데이터셋 기준 (음성 텍스트,메타정보) 통계 기반 체크	• 분석 모델 Validation Test를 통한 성능 평가 및 학습데이터 테스트

[그림 Ⅲ-34] 데이터 검수 프로세스

● 검수 단계별 세부내용

〈표 Ⅲ-10〉 데이터 검수 단계별 세부내용

검수 단계	세부 내용
작업자 검수 (수집)	• 작업자 스스로 리뷰를 진행한다.
관리자 검수 (수집/정제/가공)	• 작업자 – 관리자간의 리뷰를 진행한다. • 품질 기준에 위배되는 데이터는 재작업을 진행한다. • 범위 작업 : 데이터 수집(환경정보 수집데이터 체크) 데이터 라벨링 작업 데이터 전사화 작업 데이터 민감정보 삭제 작업
최종 검수	• 데이터 라벨링 확인 • 데이터 어노테이션 누락여부 확인 • 학습데이터 정확도 확인 • 민감정보 포함 유무 확인

● 크로스 체크를 통한 검수 일관성 유지

- 검수 단계에서검수 작업자간의 일광성을 유지하기 위해 크로스 체크를 통한 검수 진행
- 검수 작업 시작 시 별도의 교육을 통해 검수기준을 고지하고, 예시를 통한 검수기준 체득
- 검수 작업에 애매한 부분을 별도로 체크하여 정리하고, 주기적인 작업내용 및 크로스체크 사항에 대한 회의를 통해 검수 기준을 확립하고 검수 일관성 유지
- 음성 데이터의 심한 잡음에 대한 오류 체크 작업 시, 각 검수 작업자 별 검수기준이 상이할 수 있으므로, 잡음과 그 그정도에 대한 정의를 실제 음성 데이터 사례를 중심으로 교육 및 회의를 진행
- 검수 작업 중 잡음 등에 대해 검수기준이 모호한 부분 등은 해당 음성 데이터 및 내용을 공유하여 각 작업자 간의 검수 일관성 유지
- 음성 데이터에 포함된 비속어 및 은어 중 실제 자유 대화에서 사용될 수 있다고 판단되는 경우는 제외처리 하지 않고 유지
- 비속어 및 음성 검수 기준을 정의하여 검수 작업자 교육을 진행하고, 작업자 별로 비속어 및 은어 통용범위에 대해 검수 기준이 다를 수 있으므로, 회의 및 예시를 통한 검수 일관성 유지

4.2 검수 기준

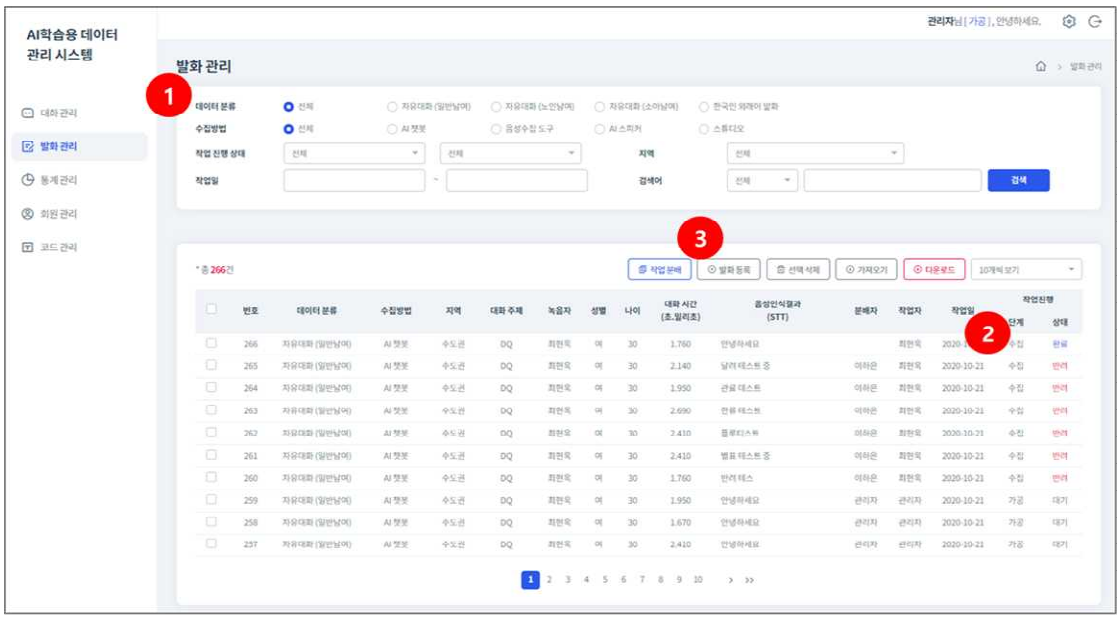
1) 검수 기준

〈표 Ⅲ-11〉 단계별 데이터 검수 기준

<table>
<tr><th>단계</th><th>검수 역할</th><th>검수 내용</th></tr>
<tr><td rowspan="2">수집</td><td>원천 데이터(음성) 수집</td><td rowspan="2">• 분류기준(카테 고리)별 음성 분류 유무 체크
• 클라핑, frame drop 등 손상된 음성신호 제외 처리 체크
• 심한 잡음으로 발생한 오류 체크
• 수집 단계 작업자의 자체 및 크로스 검증</td></tr>
<tr><td>어노테이션(수집)</td></tr>
<tr><td rowspan="4">정제 /가공</td><td>음성 데이터 가공</td><td>• 음성 데이터 텍스트화 작업</td></tr>
<tr><td>전사화 전수 검수</td><td>• 문장별 전사 데이터 일치성 체크
• 음성 파일 및 전사 텍스트 정합성 체크
• 민감 이슈(정치, 개인정보 등) 발언 저작 도구 Tool 활용하여 제외 처리</td></tr>
<tr><td>어노테이션(가공)</td><td>• 전사 데이터 라벨링 작업
• 분류 기준 별 전사 데이터 매칭 여부 체크
예) 분류 기준별 표준화에 대한 규칙 체크 및 수정
<table>
<tr><th>분류 기준</th><th>수집 데이터</th><th>표준화 데이터</th></tr>
<tr><td>성별 표준화</td><td>‘남’, ‘남자’, ‘M’ …</td><td>‘남자’</td></tr>
<tr><td>연령 표준화</td><td>‘10’, ‘10대’, ‘11’ …</td><td>‘10대’</td></tr>
</table>
• 분류 기준에 따른 어노테이션 정보(메타 정보) 매핑, 매핑 값 검수
예) 어노테이션 검증 결과 일치율 96% 이상이면 과반수에 의해 True 처리</td></tr>
<tr><td>데이터셋 검수</td><td>• 라벨링 누락 및 정합성 체크
• 문장별 전사 데이터 일치성 체크
• 전사된(음성 → 텍스트) 데이터에 대한 자체 및 크로스 검증</td></tr>
<tr><td>검수 (최종)</td><td>데이터셋 전수 검수</td><td>• 구축된 데이터셋 기반 전수 검사
예) 관리자의 통계기반 데이터 셋 체크
• 구축 데이터셋 라벨링 누락 및 정합성 체크
• 구축 데이터셋 어노테이션 누락여부 확인
• 학습데이터 정확도 확인</td></tr>
<tr><td rowspan="2">활용</td><td>Validation Test</td><td rowspan="2">• 학습 데이터 기반 서비스 확인
• 모형 성능 평가 지표 기반 분석 모델 성능 평가</td></tr>
<tr><td>학습데이터 테스트</td></tr>
</table>

2) 검수 도구 1

- 대화 관리
  - AI 음성 녹음 도구에서 자동으로 연계되어 업로드된 파일의 그룹과 관리도구의 '업로드'버튼을 통해 별도로 파일 업로드한 파일의 그룹 목록 제공
  - 데이터 분류, 대화 주제, 총 참여자, 진행상태, 등록일 등의 정보 제공 및 필터링을 통한 대화 검색 기능 제공
  - 각 데이터 그룹 별로 발화목록을 확인 및 삭제 기능 제공



[그림 Ⅲ-35] 발화 관리 도구 메뉴

① [검색조건]

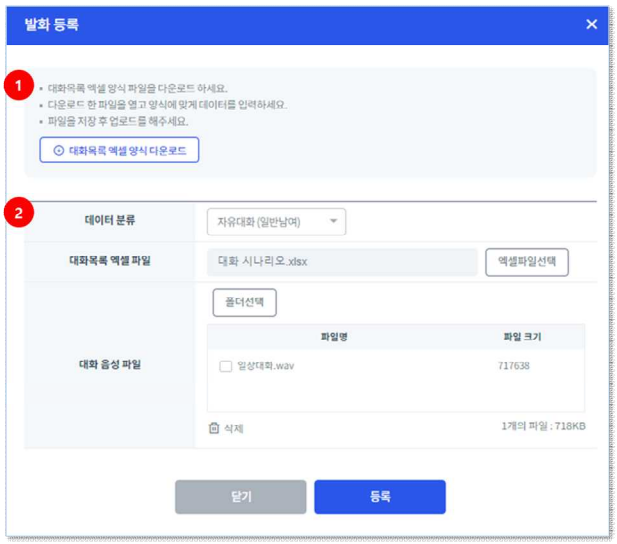
- 데이터 분류
- 자유대화(일반남여), 자유대화(노인남여), 자유대화(소아남여, 유아 등 혼합), 한국인 외래어 발화
- 진행상태
- 수집: 작업자가 대화주제에 대해 음성 녹음을 완료한 상태)
- 가공: 가공 담당자가 발화목록 검토 진행 중
- 가공 완료: 가공 담당자가 발화목록 검토 완료

- 1차 검수: 1차 검수자가 발화목록 검토 진행 중
- 1차 검수 완료: 1차 검수자가 발화목록 검토 완료
- 2차 검수: 2차 검수자가 발화목록 검토 진행 중
- 2차 검수 완료: 2차 검수자가 발화목록 검토 완료
- 검색어: 대화주제, 사용자ID

② [발화목록] 버튼 클릭하면 발화목록 관리 화면으로 이동

③ 오프라인으로 작업한 데이터를 일괄 업로드할 수 있는 화면으로 이동 (다음 화면)

- AI 데이터 파일 업로드
  - 대화 관리 화면에서 '업로드'버튼 클릭 시 레이어 팝업으로 파일 업로드 화면 실행
  - 각 데이터 분류 및 대화 주제 정의 후 대화목록 엑셀파일 및 대화 음성 파일 일괄 업로드



[그림 Ⅲ-36] 파일 업로드 화면

① 대화목록 엑셀 양식을 다운로드해서 발화목록 엑셀 작성 (엑셀 항목)

- 음성파일명
- 전사 텍스트
- 지역
- 성별

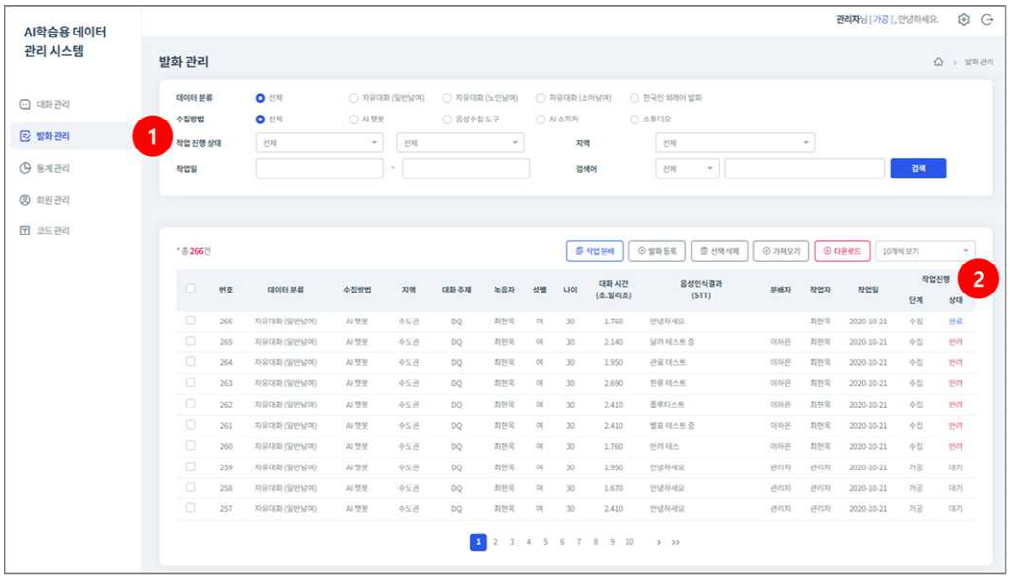
- 연령

② 데이터분류 선택

- 대화 주제 입력
- 작성한 엑셀 파일 첨부
- 대화 목록의 실제 음성파일들을 첨부

● 발화 관리

- 대화 관리 화면에서 '발화목록'버튼 클릭 시 각 데이터 그룹 내의 발화 건수 확인 및 검수 가능
- 발화내용 및 STT 일치 여부, 발화자ID, 작업 관리 진행 여부 등 확인 가능하고, 각 발화 별 '검수'버튼을 통해 검수 진행



[그림 Ⅲ-37] 발화 관리 데이터 진행상태 및 작업관리 화면

① 진행상태

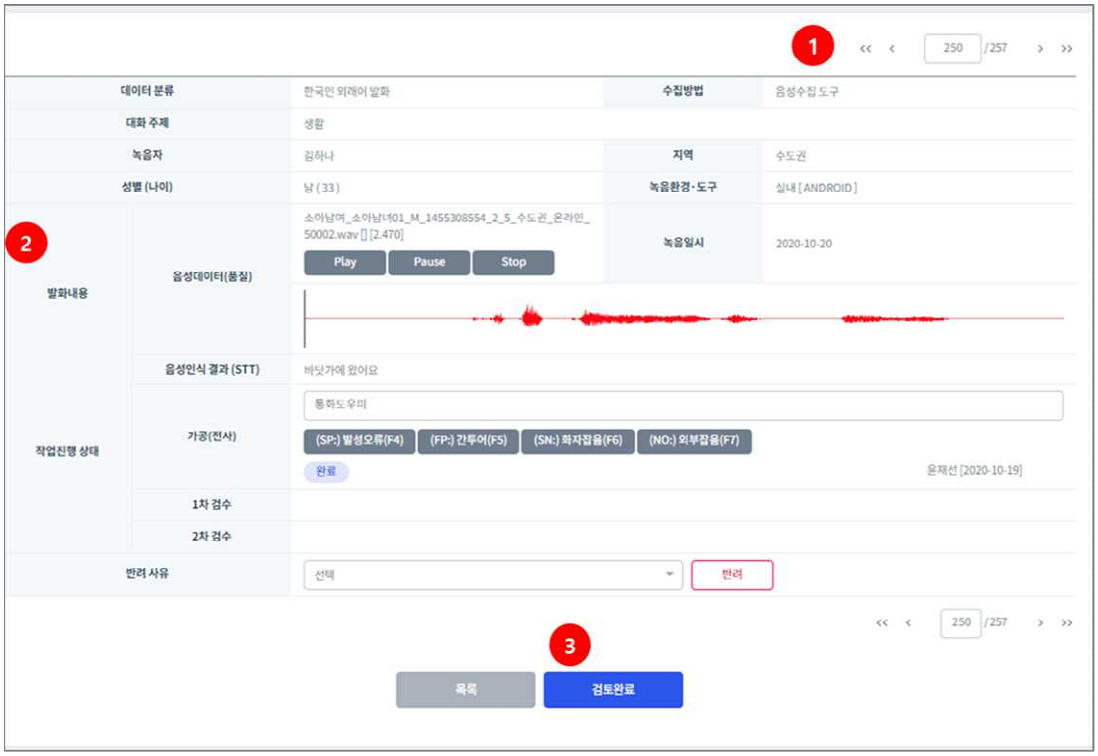
- 수집완료: 데이터 구축자가 대화주제에 대해 음성 녹음을 완료한 상태
- 가공완료: 가공자가 각 발화에 대해 가공을 완료한 상태

② 작업관리

- [검수]: 발화목록이 작업 완료인 상태로 [검수] 버튼 클릭하면 검토화면으로 이동
- 가공완료: 가공자가 검토 완료한 상태

4) 발화 검토

- 발화관리 메뉴에서 '검수'버튼 클릭 시 발화의 기본 정보(발화자ID, 지역, 성별, 연령) 확인 및 발화의 음성 데이터를 바로 재생하여 전사텍스트와 비교하여 검토 가능
- 발화 음성과 전사데이터가 동일 한 경우, 일치 선택하여 가공완료 처리하고, 불일치하는 경우 불일치 선택하여 가공완료 처리
- 일치 선택한 경우 1차 검수자 및 2차 검수자를 통해 검수 진행되고, 불일치 발화는 검수 단계로 진행되지 않음
- 검토 권한(가공, 1차 검수, 2차 검수)에 따라 발화 검토 진행



[그림 Ⅲ-38] 발화 검토 진행 화면

① 발화 순서

- [목록] 버튼을 클릭하여 발화목록 화면으로 다시 이동하지 않고, 발화검토 화면의 좌, 우 버튼 클릭으로 계속 발화 검토 가능

② 발화내용 검토

- 검토자(가공자)는 발화내용의 음성과 전사 텍스트를 각각 비교 검토하여 정확하게 일치하면 [가공 완료] 버튼을 클릭하여 검토 완료

③ 작업상태에 따라 다른 버튼명을 제공하고, 버튼 클릭 시 가공 및 검토 완료

- 작업완료: [가공 완료]
- 가공완료, 1차 검수 완료: [검토 완료]

● 검수 도구 2 : 크라우드 소싱 및 웹기반 작업 도구

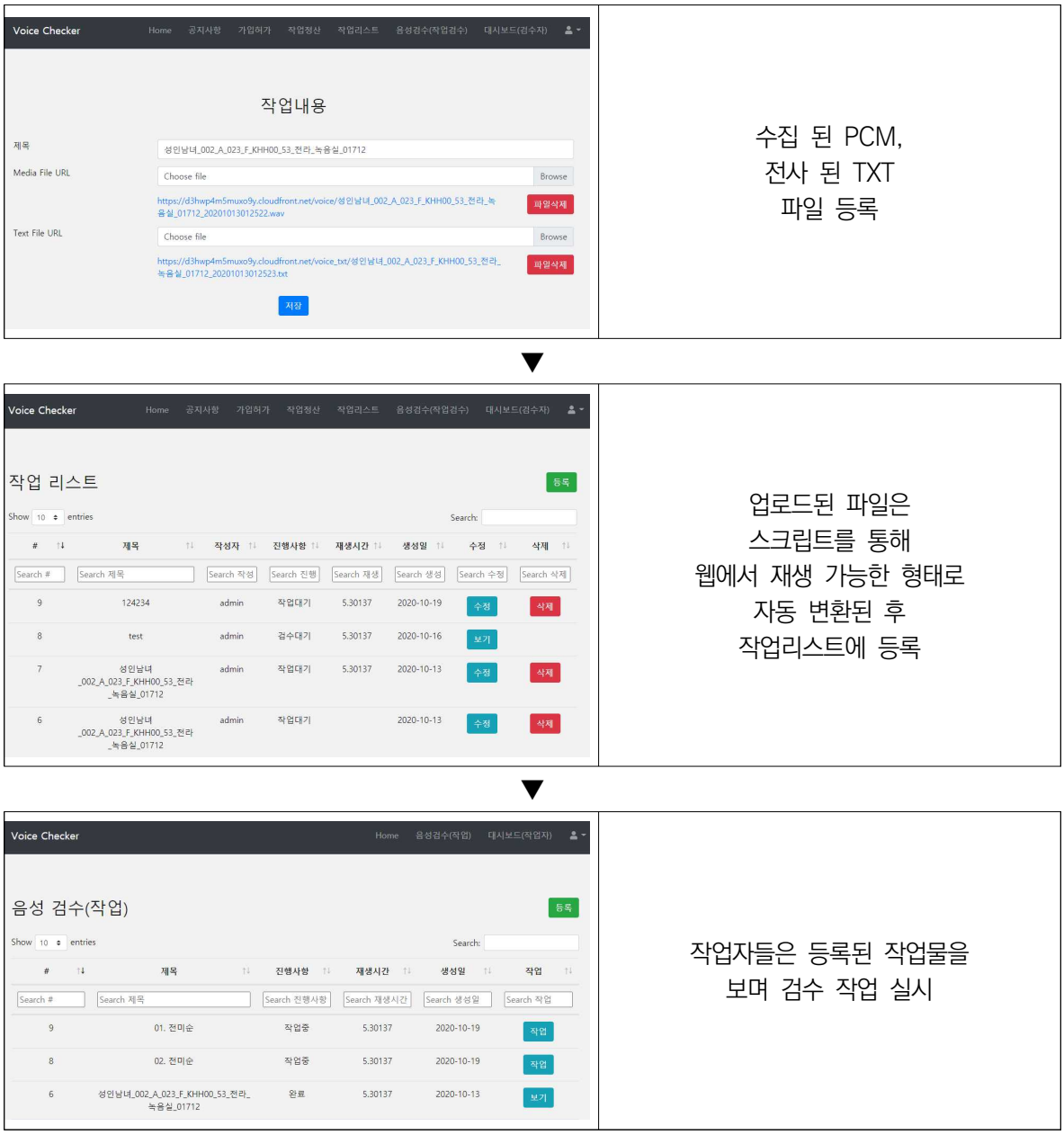
- 기본적으로 Public 크라우드소싱 / Private 크라우드소싱 작업자 모두 범용적으로 사용이 가능하며 추후 플랫폼으로써 확장성을 고려한 웹 기반 작업 도구 구축 완료

[그림 Ⅲ-39] 웹 기반 작업 도구 메인화면

- 접근성 및 편리성이 강화된 웹 기반 플랫폼을 구축하여 언제 어디서든 장소에 구애받지 않고 데이터를 검수할 수 있는 작업환경을 구축하고 작업자, 작업자를 검수하는 검수자, 총 현황을 관리하는 관리자의 3단계 유저 레벨 작업 체계를 구축하여 사업 기간 내에 과업을 완료할 수 있도록 함

관리자 (총괄 현황 관리)
▼
검수자 (작업자의 작업물 검수)
▼
작업자 (음성 전사 검수 작업)

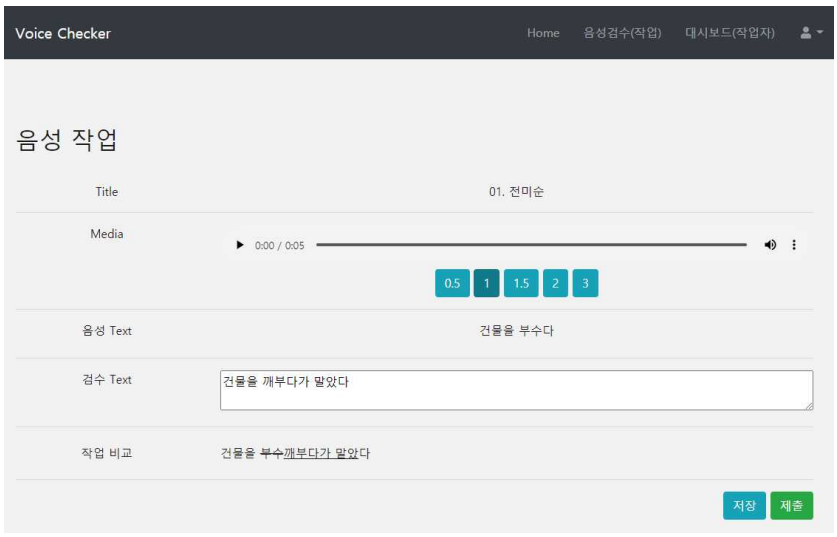
- CM, WAV, MP3 등 다양한 확장자를 가진 음성파일들을 품질 저하 없이 HTML5 웹 기반에서 구동되도록 하는 모듈 탑재



[그림 Ⅲ-40] 작업자/검수자/관리자 검수 프로세스

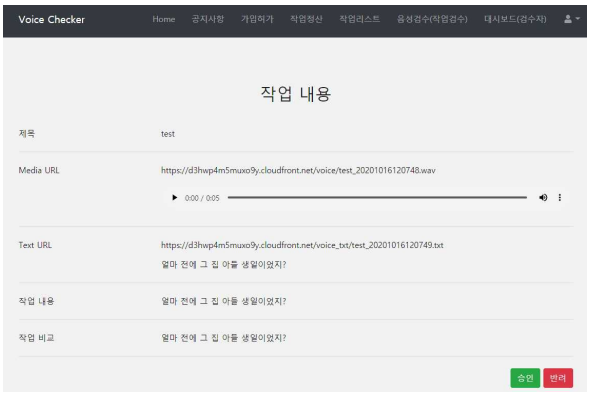
- 작업자/검수자/관리자의 검수 프로세스
  - 작업자는 1)등록된 작업을 선택하고 2)작업 화면에서 작업한 이후 3)저장 및 제출을 함으로써 1개 작업 프로세스 종료 이후 해당 프로세스 반복

- 검수자는 작업자별 제출한 작업물을 확인한 후 완료/반려 여부 결정
- 관리자는 검수자가 검수를 완료한 작업물에 대해 작업자에게 정산 실시

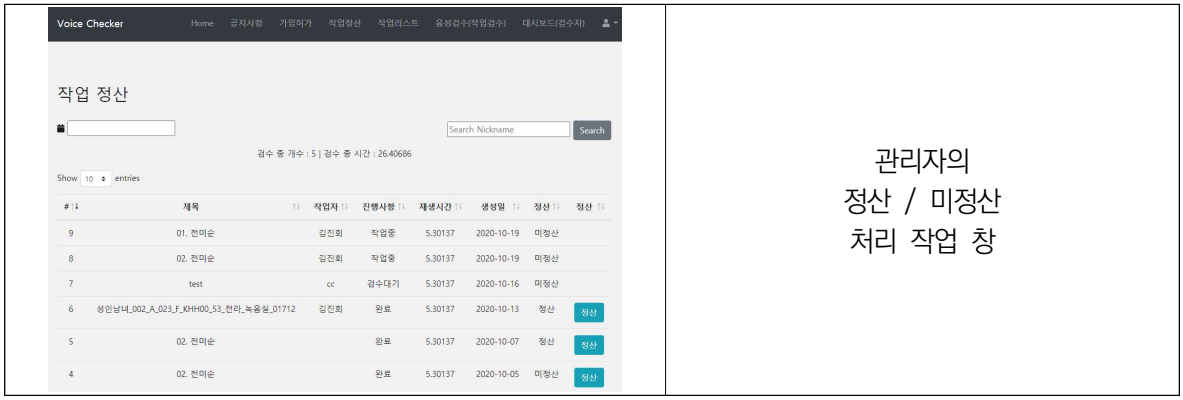


*** 작업자의 작업 화면**

- 등록된 음성을 재생하며, 전사된 파일을 보고 제대로 됐는지 검수 실시
- 검수 중 잘못 된 전사작업 확인시 수정 작업 실시
- '작업 비교'란에서 수정된 부분 자동표시



검수자의
승인 / 반려
작업 창



[그림 Ⅲ-41] 작업자/검수자/관리자 작업 화면

- 크라우드소싱 및 웹 기반 작업 도구의 장점
  - 다수의 작업자가 퍼블릭 크라우드소싱 형태로 참여가 가능하고, 가입신청 이후 승인된 작업자에 한해 작업 권한을 제공하므로, 제한된 퍼블릭 크라우드소싱 형태로 작업이 가능한 범용적인 플랫폼 형태의 도구
  - PCM, WAV, MP3 등 다양한 형태의 음성파일을 웹에서 재생될 수 있도록 자동화시키는 모듈을 탑재하여, 음성검수 작업자가 작업 진행 시 자신이 작업 내용을 '작업비교'를 통해 손쉽게 확인할 수 있는 스크립트가 짜여 있어, 활용성 및 범용성이 우수

5 데이터 활용 방안

5.1 학습 모델

- 모델학습의 의의
  - 수집된 분야별 데이터(일반남녀, 노인남녀, 소아남녀, 한국인 외래어 발화)에 대해 데이터 품질검증을 하는 방식으로 셀바스 AI BASE 모델 분야별 적응 학습을 수행하고, 인식 성능 향상됨을 확인하여 데이터의 유효성을 검증
- 데이터 1 : 자유대화(일반남녀)
  - 응용 AI 스피커로 녹음된 1,000개의 TEST DB 생성
  - 셀바스 AI BASE 인식엔진으로 음절단위 인식 성능 평가

- 데이터 1에서 수집된 다양한 환경의 학습 DB (Android, iOS, PC, AI 스피커)를 이용하여 음향 모델 학습, 서비스 도메인의 문장 기반의 언어모델 학습
- AI 스피커 음원이 포함된 적응 엔진으로 음절 단위 인식 성능 평가
- BASE / 적응 모델 간의 음성인식 성능 확인
- 적응 모델에 AI 스피커 음향모델이 적용함에 따라 인식 성능 향상 확인을 통해 수집된 데이터의 유효성 검증

● 데이터 2 : 자유대화(노인남녀)

- Android, iOS 단말로 녹음된 1,000개의 TEST DB 생성
- 셀바스 AI BASE 인식엔진으로 음절단위 인식 성능 평가
- 데이터 2 에서 수집된 다양한 환경의 학습 DB (Android, iOS, PC, AI 스피커)를 이용하여 음향 모델 학습, 서비스 도메인의 문장 기반의 언어모델 학습
- 노인 음원 음원이 포함된 적응 엔진 으로 음절 단위 인식 성능 평가
- BASE / 적응 모델 간의 음성인식 성능 확인
- 적응 모델에 노인 음원 음향모델이 적용함에 따라 인식 성능 향상 확인을 통해 수집된 데이터의 유효성 검증

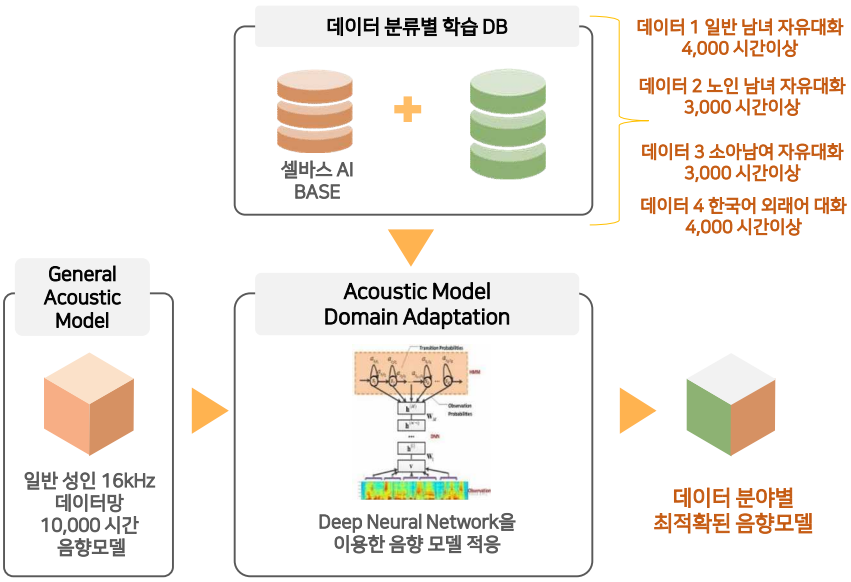
● 데이터 3 : 자유대화(소아남녀)

- Android, iOS 단말로 녹음된 1,000개의 TEST DB 생성
- 셀바스 AI BASE 인식엔진으로 음절단위 인식 성능 평가
- 데이터 3에서 수집된 다양한 환경의 학습 DB (Android, iOS, PC, AI 스피커)를 이용하여 음향 모델 학습, 서비스 도메인의 문장 기반의 언어모델 학습
- 소아 음원 음원이 포함된 적응 엔진 으로 음절 단위 인식 성능 평가
- BASE / 적응 모델 간의 음성인식 성능 확인
- 적응 모델에 소아 음원 음향모델이 적용함에 따라 인식 성능 향상 확인을 통해 수집된 데이터의 유효성 검증

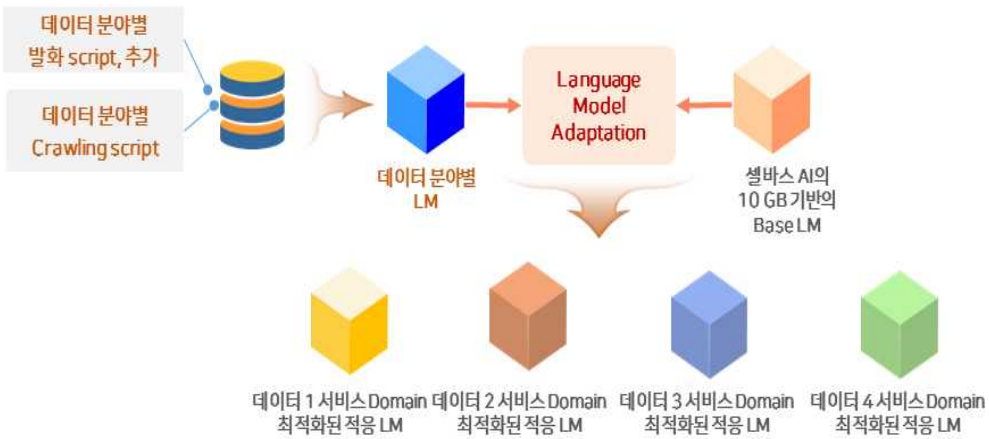
● 데이터 4 : 한국인 외래어 발화

- 외래어 가 포함된 문장으로 구성된 1,000개의 TEST DB 생성
- 셀바스 AI BASE 인식엔진으로 음절단위 인식 성능 평가

- 데이터 4에서 수집된 다양한 환경의 학습 DB (Android, iOS, PC, AI 스피커)를 이용하여 음향 모델 학습, 서비스 도메인의 문장 기반의 언어모델 학습
- 외래어 script와 음원이 포함된 적응 엔진 으로 음절 단위 인식 성능 평가
- BASE / 적응 모델 간의 음성인식 성능 확인
- 한국어 발성에 자주 나오지 않는 음소열이 포함된 외래어 음향, 언어모델이 적용함에 따라 인식 성능 향상 확인을 통해 수집된 데이터의 유효성 검증



[그림 Ⅲ-42] 데이터 분야별 최적화된 음향모델 적응학습

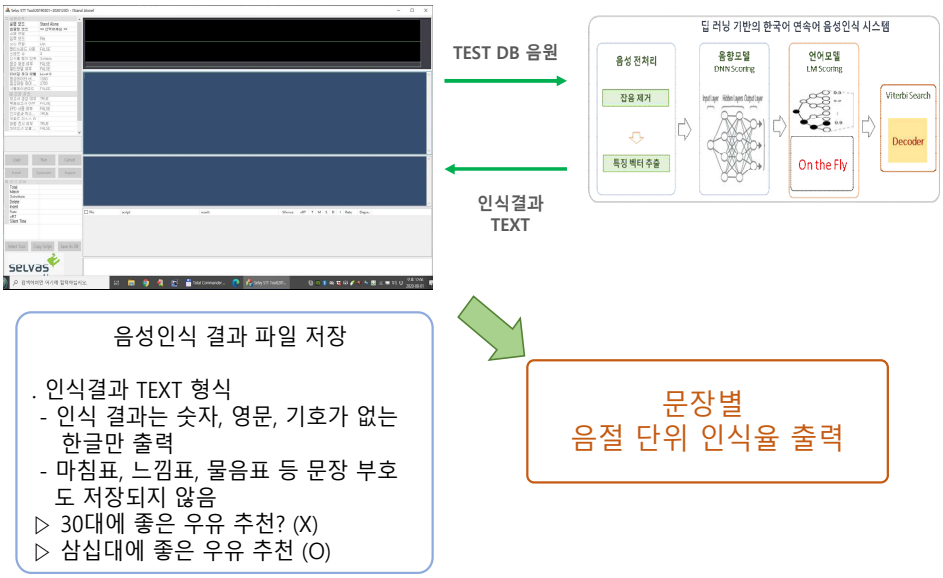


[그림 Ⅲ-43] 데이터 분야별 최적화된 음향모델 적응학습

● 한국어 음성인식 성능 평가

1) 평가 방법

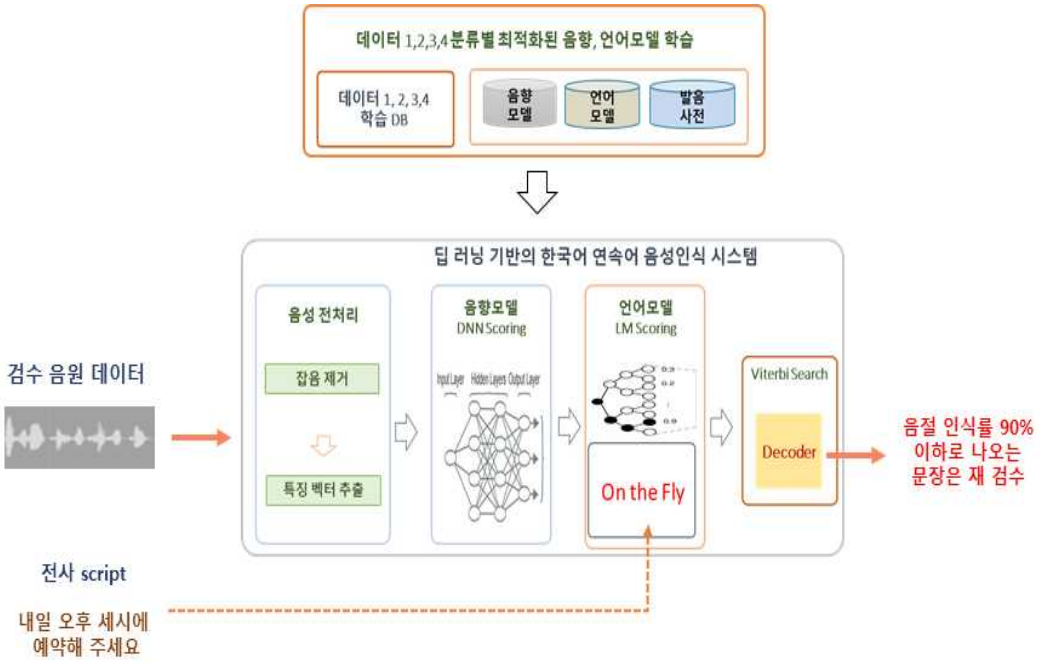
- Base 엔진, 50% AI 학습데이터 DB를 이용한 적응모델 엔진 1, 100% AI 학습데이터 적응모델 엔진 3개의 엔진별로 선별된 TEST DB에 대해 음절 단위로 비교하여 평균 인식률을 산출



[그림 Ⅲ-44] 음성인식 성능 평가

● 음성인식엔진을 이용한 최종 검수

- 전사, 검수 단계를 수행한 데이터 1,2,3,4 음성학습 DB에 대해 각 데이터별로 최적화된 인식엔진 기반으로 언어모델 On the fly 기법을 도입하여 화자별 전사 script 오류를 2차 검수 수행
- On the fly 기능은 기본 언어모델 인식 네트워크에 전사 script를 인식 후보로 올려놓고 인식을 수행함으로써 즉, On the fly에 입력된 script에 weight를 높여 인식을 하기 때문에 입력 발성과 동일한 전사 script 결과가 나오면 정상이며, 인식 결과가 다르게 나오면 전사 script에 오류가 있다고 판단
- 전사 script가 오류가 있을 경우, 음절 인식률이 90% 이하로 나오는 발화 리스트를 추출하여 재검수



[그림 Ⅲ-45] 최적화된 음성인식엔진을 이용한 최종 검수

- 최종 검수된 AI 데이터 기준으로 최종 검수를 수행하여 음절 인식률 정확도 측정 93% 확보

5.2 서비스 활용 시나리오

- 활용 시나리오

〈표 Ⅲ-12〉 서비스 활용 시나리오

과제명	AI 모델	모델 성능 지표	응용서비스
자유대화 (일반남여)	일반 남여 자유대화 모델	음절 인식율 90% 이상	챗봇 시범 서비스
자유대화 (노인남여)	노인 남여 자유대화 모델	음절 인식율 82% 이상	노인용 챗봇 서비스 “말벗”
자유대화 (소아남여, 유아 등 혼합)	소아/유아 남녀 자유대화 모델	음절 인식율 82% 이상	한국어 교육 서비스 “우리 아이 첫 한글”
한국인 외래어 발화	한국인 외래어 발화 모델	음절 인식율 82% 이상	인공지능 음악추천 서비스 “나만의 뮤직”

제4장 한국어 방언 AI 데이터

1 데이터 정보 요약

1.1 가이드 분류

대분류	오디오	중분류	자연어	소분류	WAV

1.2 데이터 정보

데이터 이름	한국어 방언 AI 데이터
활용 분야	• 연구분야: 음성 발화, 음성 인식, NLU, NLG를 포함한 NLP 전분야 • 산업분야: 온라인 심리상담, 고객상담 챗봇, 스마트 스피커 등
데이터 요약	• 한국어 방언 데이터 구성(안) : 한국어 방언(강원도, 경상도, 전라도, 제주도, 충청도)의 총 250만 문장의 학습데이터

1.3 데이터 구축 개요

- AI 학습용 데이터 구축량 : 한국어 방언(강원도, 경상도, 전라도, 제주도, 충청도)의 총 250만 문장의 99.9% 고품질 학습데이터 구축 및 AI 응용서비스 개발
- 데이터 구축 프로세스는 한국어 방언 인공지능 학습용 데이터 구축 → 데이터 품질 관리 및 검증방안 → AI 데이터 활용 응용서비스 개발로 진행

(고품질 학습데이터) 한국어 방언 (5개도) 구축 총 3,000시간의 음성 데이터 수집 및 50만 문장전사	• 원천 데이터 수집 시 정제 (사전녹음, 녹음수행, 완료파일청취 등) • 학습데이터 설계 및 대화 주제 선정 (표준어와 방언 매핑) • 음성 녹음 화자 구성 및 절차 수립 (각 세부과제 별 인구통계기준) • 음성 수집 비율과 수집 도구 (대면/비대면, 녹음도구/화상 녹취)
(학습데이터 품질관리) 4단계 품질 공정 및 도구	• 고품질 학습용 데이터 확보를 위한 검수 방안 제시 • 참여기업 및 세부책임, 수행기관, 품질기관 검수 → 99.9 %품질 • 음성전사 저작도구 활용한 투입작업 인력 품질관리 • 음성 녹음 지원자 모집 및 수집 평가 (크라우드소싱, 데이터품질평가)
(AI 응용서비스) 서비스 적용사례 4가지 개발	• 각 도별 5개+ 통합 1개 (음성인식,합성,기계번역,일상대화모델) • 4가지 서비스 적용사례 제시 • 솔트룩스 AI클라우드 3년간 무상제공 → 학습된 AI모델 활용

[그림 Ⅲ-46] 데이터 구축 개요

1.4 구축 목적

- 데이터 경제로의 패러다임 변화
  - 4차 산업혁명 시대로 급속 진입하면서 제조업, 서비스업 중심의 한국경제는 도태의 위기에 직면하게 됨. 특히 코로나19로 인한 극심한 경기 침체와 함께 데이터 경제로의 패러다임 전환이라는 이중 과제를 해결해야 하는 시점
  - 데이터를 기반으로 한 인공지능의 시대가 도래함 에 따라 인공지능 시대의 석유라고 일컫는 기초 데이터의 국적 차원의 확보 및 제공이 글로벌적인 경쟁력 확보의 필수 요소이며, 데이터 확보가 이루어져야 비로소 디지털 시대로의 전환기를 맡은 수많은 기업과 스타트업 그리고 국가 공공 행정 서비스의 미래 선도형 경제 실현이 가능한 시점에 도달
  - 이미 다른 선진국에서는 미래 경쟁력을 좌우하는 데이터의 중요성을 인식, 데이터 산업 활성화를 위해 국가 차원의 선제적인 전략 수립과 정책 투자 확대 등 데이터 경쟁에 돌입
  - 구글, 아마존 등 글로벌 IT 대기업은 빅데이터의 축적과 함께 다양한 AI 혁신기술을 공개하며 수많은 형태의 새로운 산업과 서비스 영역을 개척하며 선보이고 있어 벌써부터“디지털 독과점”이란 비판을 받고 있는 수준으로 앞서 나가고 있음
- 한국어 방언 데이터가 필요한 이유
  - 모든 디지털 산업의 기초가 될 데이터는 80% 이상이 텍스트, 음성, 영상 등으로 되어 있음. 이중 음성, 텍스트 데이터는 인공지능 A.I를 학습시키기 위한 기술인 NLP(Natural Language Processing)의 핵심적인 부분을 차지함

- 그러나 한국어 말뭉치를 비롯한 원천 데이터 구축은 선진국과 글로벌기업 대비 걸음마 수준이며, 관련 업체 수요에도 부응하지 못하는 열악한 수준임. 따라서 디지털 뉴딜에서의 한국어와 한국어 방언 데이터의 수집 및 구축 사업은 인공지능 학습용 데이터 구축 사업의 가장 근간이 되는 중요한 부분이라 할 수 있음

1.5 활용 분야

- 연구분야: 음성 발화, 음성 인식, NLU, NLG를 포함한 NLP 전분야
  산업분야: 온라인 심리상담, 고객상담 챗봇, 스마트 스피커 등

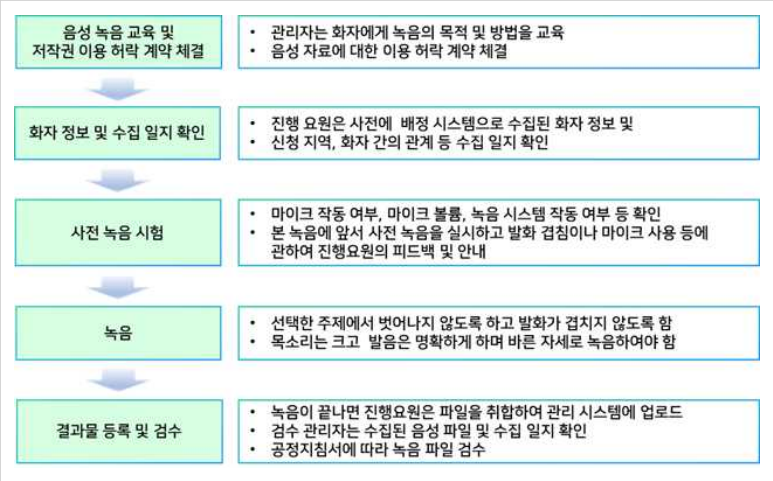
1.6 유의 사항

- 저작권 이용 허락 계약
  - 관리자는 화자에게 사업 설명 및 녹음의 목적을 설명하고 저작권 이용 허락 동의서를 체결
- 동의서 체결 프로세스화
  - 동의서 체결이 되지 않을 경우는 화자에 참여 시키지 않음

한국어 방언 AI 데이터 구축 및 활용 저작권 이용허락 동의서

한국정보화진흥원은 대규모 언어 자원을 구축하여, 이를 국어 연구 및 자연언어처리 기술 개발 등을 위해 사용하고자 합니다. 이에 귀하의 저작물의 활용에 대한 승인을 구하고자 합니다. 귀하가 만드신 저작물은 개별 단어 및 문장, 텍스트에 대한 정보 추출과 분석에 쓰이며, 귀하의 저작물을 이용하여 구축한 방언 데이터가 귀하의 저작·출판에 관한 어떠한 권리에도 손상을 입히지 않을 것임을 약속드립니다. 모쪼록 21세기 4차 산업 혁명 시대에 귀하의 저작물을 우리말 정보 처리 발전의 기초가 되는 국가적 언어 자원의 구축에 유용하게 활용할 수 있도록 아래와 같이 협조하여 주시면 감사하겠습니다.

[그림 Ⅲ-47] 저작권 이용허락 동의서



[그림 Ⅲ-48] 녹음 데이터 수집 프로세스

2 데이터 획득 및 정제

2.1 원시 데이터 선정

- 녹음 환경 및 화자 모집
  - 녹음 품질 및 데이터 표본의 다양화를 위해 녹음지역을 전국에 있는 지사 사무실을 활용하여 진행하며, 두 명의 화자가 서로 자유롭고 편안한 상황에서 대화할 수 있도록 1 조 2 인 단위로 화자를 모집하여 자료 수집을 진행
- 녹음 환경 구성
  - 두 명의 화자가 편안하게 이야기 할 수 있는 사무실 환경 마련
  - 녹음실은 외부와 차단된 상태로 대화에 참여한 두 명만이 대화할 수 있도록 구성
  - 화자는 각각 헤드셋 마이크를 착용하고 발화
  - 상대방의 목소리가 들어가지 않도록 적정거리 유지
- 녹음 화자 모집
  - 특정 성별, 연령, 지역 등이 편중되지 않도록 사전 협의하여 진행
  - 한 화자당 최대 녹음시간은 가능한 약 30분으로 하고 동일 화자가 중복 참여하지 않도록 제한하나, 동일 주제가 아닐 경우에는 허용

- 녹음 화자 모집 시 최초 2인 1조로 신청자를 최우선으로 하며, 1인이 개별 신청했을 경우 비슷한 연령대 및 관심사를 구분하여 조 편성
- 주제에 따라 1인 녹음, 3인 이상 녹음을 허용

〈표 Ⅲ-13〉 화자별 분류 기준

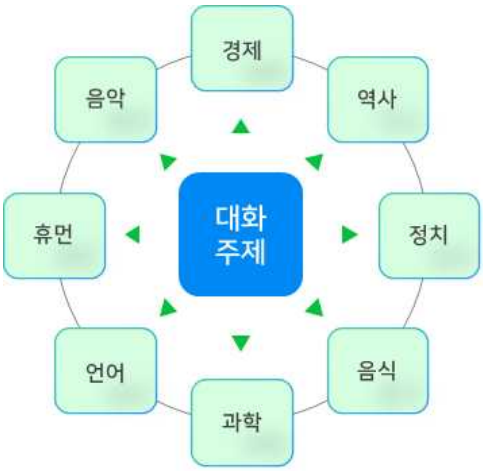
화자별 분류방법	세부 내용
연령별	1그룹(10대~20대), 2그룹(30~40대), 3그룹(50대 이상)
지역별	강원도, 경상도, 전라도, 제주도, 충청도

● 녹음 화자 구성

- 연령별 그룹 기준으로 분류
- 총 3개의 그룹으로 구성되며 1그룹은 10대~20대, 2그룹은 30~40대, 3그룹은 50대 이상으로 정하였으며 실제 녹음이 불가능하다고 판단되는 0~9세 / 70세 이상의 대상자는 제외이나 녹음이 가능할 경우 3그룹에 포함하여 진행
- 최종 인력 배분은 NIA(한국정보화진흥원) 협의 후 진행
- 화자 비율은 1그룹 40%, 2그룹 20%, 3그룹을 20%를 최소 비율로 하고 최대 비율은 1그룹은 45%, 2그룹과 3그룹은 합하여 15%를 구축함
- 단, 2그룹과 3그룹의 편차는 5% 이상을 넘지 않게 함

[그림 Ⅲ-49] 연령별 화자 비율 구분

- 대화 주제 분류
  - 녹음 화자가 편중된 내용을 발언하지 않도록 체계적인 주제와 자료를 제시하여 화자가 원활하고 편안한 환경에서 대화를 나눌 수 있도록 하며, 녹음 상황임을 인지하지 않고 자연스럽게 참여할 수 있도록 환경 조성
  - 대화 주제 목록 (예시)



[그림 Ⅲ-50] 대화 주제 제시 자료 - 사진, 그림, 멀티미디어 등 (예시)

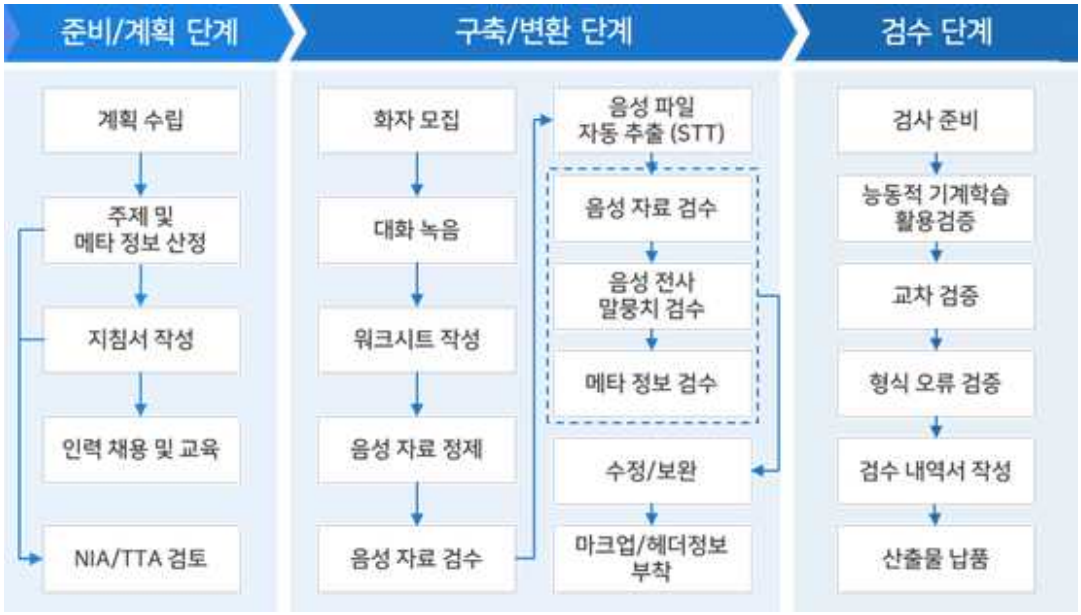
2.2 규제관련 사항

- 관련 내용 없음

※ 본 데이터 가이드라인 상에는 규제 관련 내용이 없지만, 데이터 구축 시 규제 관련 내용 확인 및 검토 필요

2.3 획득 및 정제 절차

- 한국정보화진흥원의 데이터베이스 구축방법론(Ver.4)을 적용하여 음성 녹음, 이중 전사, 원시 말뭉치 구축에 대한 대상 자료별 공정 태스크와 주요 활동 절차를 표준화하여 효율적인 학습용 데이터셋 구축 체계를 확보하고 한국어 방언 AI데이터 구축에 적합하도록 자료의 특성을 고려하여 준비/계획단계, 구축/변환단계, 검수단계의 3단계 공정을 설계한다.



[그림 Ⅲ-51] 획득 및 정제를 위한 3단계 공정

2.4 획득 및 정제 기준

- 관련 내용 없음

3 어노테이션/라벨링

3.1 어노테이션 / 라벨링 절차

가. 준비/계획단계

1) 학습데이터 구축 대상 및 범위 계획 수립

● 말뭉치 구축 대상 및 범위 선정

- 두 사람이 특정주제 (10개 내외)로 자유롭게 대화
- 대화 내용을 녹음하고 정제
  (2,000명 이상의 화자가 발화한 총 3,000시간 이상 데이터셋, 대화당 15분 이하)
- 해당 녹음자료에 대한 저작권 이용 허락 계약서 체결
- 녹음된 내용 이중 전사 (발음전사 / 철자전사)

- 구축된 전사자료에 대한 메타정보 구축 (녹음날짜, 화자명 및 정보, 대화주제 등)

세부 메타정보 항목	날짜, 대화명, 주제, 화자명, 화자정보 , 화자간 관계 등 ※ 단, 화자 개인정보는 비식별화
	예) 1. 녹음날짜 - 2020년 3월 26일 2. 참 여 자 - ① 김철수 (33세, 남, 직장인, 부산출생, 현 서울거주) ② 최영희 (31세, 여, 직장인, 서울출생, 현 서울거주) 3. 대화자관계 – 회사 거래처 담당자 4. 대화주제 – 주말 일상

[그림 Ⅲ-52] 전사자료에 대한 메타정보 항목

● 학습데이터 주제 선정

- 컨소시엄의 인공지능 전문가 그룹을 통해 방언 학습데이터 구축을 위한 상세 분석 후 주제 및 메타 정보를 검토
- NIA(한국정보화진흥원) 검토 및 협의 후 최종 주제와 메타를 선정

[그림 Ⅲ-53] 학습데이터 주제 선정 과정

나. 구축/변환단계

1) 학습데이터 구축 및 변환 작업

● 녹음 환경 및 화자 모집

- 녹음 품질 및 데이터 표본의 다양화를 위해 녹음지역을 전국에 있는 지사 사무실을 활용하여 진행하며, 두 명의 화자가 서로 자유롭고 편안한 상황에서 대화할 수 있도록 1 조 2 인 단위로 화자를 모집하여 자료 수집을 진행

● 음성 녹음 및 정제

- 음성 녹음된 자료가 기준에 벗어나지 않도록 녹음 시 국립국어원과 협의된 기준에 부합하도록 작업하며, 녹음된 음성 자료를 전사 단위로 편집하고 개인 정보 및 불필요한 내용을 정제하는 과정도 효율적인 체계를 구축하여 업무상 인적, 물적 손실을 최소화 합니다.
- 대화 전체 음성 파일(원본) / 억양구 단위로 분할된 파일(정제본)을 각각 제출

※ 단위 등 정제 기준은 주관기관과 사전 협의

※ 폴더 구조 및 파일명 등은 주관기관 제시자료를 따르되 상세 내용은 사전 협의

- 대화 주제와 무관한 내용(인사말 등)은 제외하여 정제

● 전사 및 학습데이터 구축

- 화자를 통해 녹음 및 정제가 된 음성 자료를 대상으로 작업 지침에 따라 전사 작업을 수행하고 학습 데이터의 정확성 확보를 위해 교차 검수를 진행하여 100% 정확도를 기함
- 전사 지침

※ 발화된 그대로 전사하는 발음전사와 한글 맞춤법 및 표준어 규정에 따른 철자 전사를 병행하여 전사하는 것을 기본 원칙으로 함

※ 그 외 화자표시, 전사단위, 발화겹침, 불완전한 발화, 띄어쓰기 등 세부 내용은 국립국어원이 제시한 전사지침을 따름

- 학습데이터 지침

※ 전사 결과물에 대해 헤더정보 부착 등 표지 부착 작업 수행

※ 파일명 부여방식, 표지 부착방식, 형식 등은 협의하여 진행

〈표 Ⅲ-14〉 전사 및 학습데이터 구축 과정

절차	내 용	담당	산출물
음성 전사 학습데이터 교정	• 작업지침 교육 및 전사 프로그램 교육 • 음성 텍스트 자동 변환(STT) 후 전사 학습데이터 교정 • 발음 전사와 철자 전사가 병행 전사 • 영문과 숫자 한글로 표기되었는지 점검 • UTF-8로 저장	구축팀	워크시트 작업지침서
메타 정보 점검	• 메타 정보가 일치하는지 확인 • 메타 정보의 오탈자 및 중복 확인 • 불일치 데이터 재작업	구축팀	워크시트 작업지침서

절차	내 용	담당	산출물
전사 학습데이터 마크업	• 점검이 완료된 전사 학습데이터 마크업 • 표준 지침에 따라 변환 • 자동 변환된 마크업 점검	구축팀	작업지침서 점검내역서
인계	• 점검이 완료된 음성 자료와 음성 전사 학습데이터/메타정보는 품질 점검팀으로 인계	구축팀	음성 자료 음성 전사 학습데이터

3.2 어노테이션 / 라벨링 기준

〈표 Ⅲ-15〉 음성 전사 규칙

분류	전사규칙
개요	• 발화된 그대로 전사하는 발음 전사와 한글 맞춤법 및 표준어 규정에 따른 철자 전사를 병행하여 전사하는 것을 기본 원칙으로 한다.
화자 표시	• 화자 아이디, 성별, 연령, 직업, 출생지, 주 성장지, 현 거주지, 학력 등 화자 정보를 표시한다. 화자에 대한 정보를 모를 경우에는 '?'로 표시한다. • 본문 전사에서 화자 정보와 화자 표시는 반드시 일치해야 하고 화자가 분명하지 않을 경우에는 '?'로 표시한다.
전사 단위	• 기본 전사 단위는 긴 휴지, 경계 억양, 경계말 장음화 등을 특징으로 하는 억양구(IP: Intonational Phrase)가 되도록 한다.
숫자/외래어/기호/단위	• 한글 맞춤법, 표준어 규정, 외래어 표기법 등 관련 어문 규정에 따라 한글로 적는다.
발음	• 표준 발음에서 벗어난 형식으로 발화하거나 표준 발음이 여러 개인 경우 등 한글 맞춤법 표기에 따른 발음과 차이가 있는 경우에 발음 전사를 한다.
발화 겹침	• 겹침 발화는 표시하지 않고 시간 순서에 따라 적는다. • 만약 맞장구 발화가 일어날 경우 맞장구 발화를 사이에 넣어 주 발화를 나눈다.
익명성 보장	• 대화자들의 신분 보장을 위해 이름, 주민등록번호, 카드 번호, 전화 번호 등 개인 정보와 관련된 사항은 노출되지 않도록 전사 단계에서 비식별화한다. - n : 사람 이름(단, 정치인, 연예인 등 유명인의 이름은 비식별화하지 않으며, 상호명은 부정적인 경우에만 비식별화) - social-security-num : 주민등록번호 - card-num : 신용카드 번호 - address : 주소(동 이하의 구체적인 주소만 비식별화) - tel-num : 전화 번호
기타 소리	• 기타 소리 중 웃음, 목청, 박수, 노래에 대한 4가지는 태그로만 전사한다. • 기침, 들숨, 날숨, 재채기, 코훌쩍, 하품 등은 전사하지 않는다.

분류	전사규칙
축약형 표기	• 언어 경제성의 원칙에 의해 구어에서는 축약형이 많이 나타나며, 이는 모두 표기에 반영 • 모음 축약형은 ’를 사용해서 두 음소를 연결해 준다.

3.3 어노테이션 / 라벨링 교육

- 전사 작업자 대상 작업 지침 교육
  - 전사 작업장 내 마련되어있는 교육장을 활용하여 대상자 교육 실시
  - 전사 관련 기본교육은 상시 실시하고 있으며, 본 사업에서 국립국어원이 제시한 규정을 재확인하여 해당 규정에 맞는 교육 실시
- 전사 프로그램 교육

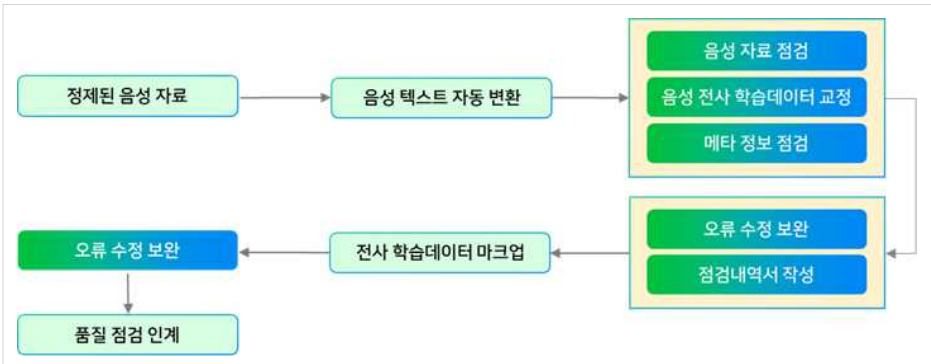
3.4 어노테이션 / 라벨링 도구 및 사용법

- 관련 내용 없음

4 데이터 검수

4.1 검수 절차

- “방언 전사”와 “표준어 대응쌍”은 “(방언 전사 형태)/(표준어 대응쌍 형태)”의 꼴로 제시하고, 방언 전사한 내용이 표준어와 차이가 없을 경우는 방언 전사의 내용을 그대로 유지한다.



[그림 Ⅲ-54] 음성 전사 검수 절차

4.2 검수 기준

● 기본 원칙

1) 이 작업은 한국어 방언 AI데이터 구축을 위하여 5개의 지역으로 묶어서 수집한 방언을 일차적으로는 지역 방언의 특성을 살려 "방언 전사"하고, 표준어 규정에서 벗어나는 방언에 해당하는 부분에 대해 "표준어 대응쌍"을 제시하는 것이다.

2) “방언 전사”와 “표준어 대응쌍”은 “(방언 전사 형태)/(표준어 대응쌍 형태)”의 꼴로 제시하고, 방언 전사한 내용이 표준어와 차이가 없을 경우는 방언 전사의 내용을 그대로 유지한다.

〈표 Ⅲ-16〉 방언 전사 내용

지역	보기
강원	이게 (다나?)/(다니?) 나도 이쪽 동네 (출신이래.)/(출신이야.) (이라)/(이렇게)
경상	어제 어디 (갔었노?)/(갔었니?) 미역 (쭐거리)/(줄기) (단디)/(단단히)
전라	혼자 다 (묵어)/(먹어) (분당께.)/(버린다니까.) 아 (실맹키로)/(실처럼) 가는 거 그거? (그랑께)/(그러니까)
제주	아까 (집드레)/(집으로) (가라.)/(가더라.) 너 (하*구정 한)/(하고 싶은) 대로 (하*라.)/(해라.) (아매나)/(아무렇게나)
충청	동네 사람들은 (워떡헌댜?)/(어떡한대?) 가만히 (두덜)/(두질) (못햐.)/(못해.) (그려.)/(그래.)

※ [참고] 제주 방언에 제시된 ‘하*’에서 ‘아*’는 아래아(·)를 나타내는 표기이다. ‘3.13. 지역별 방언 전사 주의 사항’ 참조(17쪽)

3) 방언 전사하기

- ① 방언 전사: 각 지역에서 모은 사람들의 대화를 지역 언어의 특성이 드러나도록 소리나는 대로 적는 것.
- ② 방법 및 유의 사항: 방언과 관련이 없는 표현은 표준어를 적는 방식으로 쓰되, 방언 표현은 방언의 특색을 드러나도록 표기한다. 이때 방언의 표기는 음성 그대로 소리나는 대로 쓰지 않고 방언의 형태가 드러나는 방식으로 쓴다. 예를 들어 “먹었지”의 경상도 방언은 [묵었지]로 쓰며, 소리나는 대로 연음한 [무걷찌]로 쓰지 않는다.

예) 올바른 전사 표기: (묵었지)/(먹었지), (갔었노?)/(갔었니?)
잘못된 전사 표기: (무걷찌)/(먹었지), (가썬노?)/(갔었니?)

〈표 Ⅲ-17〉 방언 전사 예시

지역	보기	
	올바른 전사 표기	잘못된 전사 표기
강원	그때는 (우뚱게)/(어떻게) 할 수가 없었어.	그때는 (우뜨케)/(어떻게) 할 수가 없었어.
경상	어제 어디 (갔었노?)/(갔었니?) 오늘 날씨가 너무 (추버서)/(추워서)	어제 어디 (갇었노?)/(갔었니?) 어제 어디 (가썬노?)/(갔었니?) 오늘 날씨가 너무 (춥어서)/(추워서)
전라	니가 (멋이간디)/(뭣이관데) 큰 소리냐?	니가 (머시간디)/(뭣이관데) 큰 소리냐?
제주	하루에 같이 (검질멧주게.)/(김매었지).	하루에 같이 (검질멧쭈게.)/(김매었지). 하루에 같이 (검질멛주게.)/(김매었지). 하루에 같이 (검질멛쭈게.)/(김매었지).
충청	돈은 물어 주면 되지만 속상한 건 (위칙헌댜?)/(어떡한대?)	돈은 물어 주면 되지만 속상한 건 (위치컨댜?)/(어떡한대?) 돈은 물어 주면 되지만 속상한 건 (위치껀댜?)/(어떡한대?)

4) “표준어 대응쌍 전사”는 소리 나는 대로 적은 “방언 전사”가 표준어 규정에서 벗어난 경우에, 그에 대응하는 표준형을 함께 제시하는 것을 원칙으로 한다. 띄어쓰기를 기준으로 하여 방언과 표준어를 각각 괄호 안에 넣어서 전사하고 이들 사이에는 빗금(/)을 넣는다. 방언 전사를 먼저하고 표준어 대응쌍 전사를 그 뒤에 나란히 제시한다.

〈표 Ⅲ-18〉 방언 전사-표준어 대응쌍 전사 예시

지역	보기
강원	(여서)/(여기서) 꾸물거리지 말고 (얼푼)/(얼른) 가라. 마을 사람들은 뭐든 (농가)/(나누어) 먹지요.
경상	근데 (지)/(자기) 생각이 옛날부터 그런 생각을 하더라고. 여기에 동그라미나 (꼽표)/(곱표) 치세요.
전라	아이들은 (훈지)/(그네) 뛰면서 놀고 있었다. 늦은 사람이 (댑대로)/(도리어) 큰소리친다.
제주	성격이 참 (요망지다.)/(야무지다.) (하르방)/(할아버지) 댁에 가는 길.
충청	여기 (졸)/(부추) 한 단에 얼마요? (고쿠락)/(아궁이) 불이 꺼졌나 좀 봐라.

〈표 Ⅲ-19〉 방언 전사-표준어 대응쌍 전사 예시

지역	보기
강원	모처럼 해가 난 (날에느)/(날에는) 마실이나 (댕게오시우.)/(다녀오시오.) 애가 종일 (울민서)/(울면서) 쳐다보더라고. 돈이 (읎어도)/(없어도) 남한테 (아수운)/(아쉬운) 소리는 못하겠다.
경상	상담소에는 어떤 걸 기대하고 (왔으까?)/(왔을까?) 마음에 든다 싶으면은 그냥 다 하는 (스타일이래)/(스타일이어) 가지고. (글잖아)/(그렇잖아). (우리가잉)/(우리가) 그래도 둘은 됐으면 하는 생각이 있잖아.
전라	하루 종일 이엉만 (영끄고)/(엮고) 급히 약을 지었는데도 못 (나수고)/(낫고) (가 부렀어.)/(가 버렸어.) 그거 다 (이애기헐라면)/(이야기하려면) (미칠을)/(며칠을) 해도 안 돼.
제주	(게민)/(그러면) (모지레민)/(모자르면) (멋을)/(뭣을) 더 (가*라 주코?)/(말해 줄까?) 야 (무신)/(무슨) 그런 게 또 (시어.)/(있어.) 어떻든 (저디)/(저기) 다 (지내치민)/(지나치면) (되우다.)/(됩니다.)
충청	너 (또래)/(때문에) (여기꺼지)/(여기까지) 와야 (되겄어?)/(되겠어?) (오둥아를)/(오디를) 얼마나 (마이)/(많이) 먹었는지 입 안이 시커멓게 (물들었슈.)/(물들었어요.) 점심 때는 (밥얼)/(밥을) 먹구, (새이)/(새참) 때는 (국시를)/(국수를) 먹는 (겨)/(거야).

5) 외래어, 외국어의 경우 들리는 대로 한글로 표기하고 대응쌍은 전사하지 않는다. 이 사업은 방언형의 표준형 대응쌍 이중전사를 정확하게 하는 것이 목적이다. 따라서 작업 효율을 고려하여 외래어, 외국어는 사전 표준형을 일일이 검색하기보다는 들리는 대로 적도록 한다. 예를 들어, 발화자가 '뻐쓰'로 발화한다면 '뻐쓰'만 전사하면 된다.

〈표 Ⅲ-20〉 방언 전사-표준어 대응쌍 전사 예시

올바른 표기	잘못된 표기
아무래도 뻐쓰가 더 빠르지. ([뻐쓰]로 발음한 경우)	아무래도 버스가 더 빠르지. 아무래도 (뻐쓰가)/(버스가) 더 빠르지.
그린 뉴딜 시대에 맞는 그린 모빌리티 보급 확대	Green-New Deal 시대에 맞는 green mobility 보급 확대

6) 숫자, 외국어, 기호, 단위 등은 숫자나 기호가 아닌 한글로 표기한다. 규범 표기가 미확정된 외국어의 경우 〈우리말샘〉의 등재된 표기를 기준으로 삼는다.

〈표 Ⅲ-21〉 방언 전사-표준어 대응쌍 전사 예시

올바른 표기	잘못된 표기
우리가 구 박 십 일 갔었나?	우리가 9박 10일 갔었나?
시급이 오천 원 육천 원짜리도 하는 데도	시급이 5천 원 6천 원짜리도 하는 데도 시급이 5000원 6000원짜리도 하는 데도
그거는 항상 백 프로 만족은 잘 없다라고 생각이 듭니다.	그거는 항상 100% 만족은 잘 없다라고 생각이 듭니다.
그린 뉴딜 시대에 맞는 그린 모빌리티 보급 확대	Green-New Deal 시대에 맞는 green mobility 보급 확대

- 화자 표시
  - 화자 아이디, 성별, 연령, 직업, 출생지, 주 성장지, 현 거주지, 학력 등 화자 정보를 표시한다. 화자에 대한 정보를 모를 경우에는 'NA'로 표시한다.
  - 본문 전사에서 화자 정보와 화자 표시는 반드시 일치해야 하고 화자가 분명하지 않을 경우에는 'NA'로 표시한다.
- 방언 전사의 단위 구획과 문장 부호의 사용

1) 방언 전사의 단위 구분

- 방언 전사의 단위는 문장 단위를 기본 단위로 하되, 구어의 특성을 고려하여 너무 긴 문장은 쉼을 고려하여 단위를 구분한다.
- 글을 쓸 때에는 마침표로 문장이 끝났음을 알 수 있지만, 방언에서는 문장부호가 없으므로 문장과 유사한 단위를 기본 단위로 구획할 필요가 있다. 이에 대한 기준은 다음과 같다.

① 문장 단위를 기본 단위로 한다. 단 문장이 너무 길거나 중간에 쉼이 있는 경우는 아래 ②와 같이 단위를 구분한다.

② 한 단위로 전사하는 분량이 6초를 넘지 않도록 제한하며, 이때 띄어쓰기 단위는 10개 이내가 된다. 이 기준은 절대적인 기준은 아니며, 컴퓨터의 자동 의미 분석에서 단위가 지나치게 복잡해지도록 하지 않기 위함이다. 즉 말하는 사람의 나이나 말하는 속도에 따라 6초 이상이 될 수도 있다.

③ 단위의 구분은 문장 내용의 달라짐을 기준으로 한다. 가급적 하나의 문장이 완성될 때에 단위를 나눈다.

예) 전사 단위 구획 예(※음성파일 sample01.pcm 참고)
※ 총 7구간. 각 구간마다 5~6초로 나눔

2) 문장 부호의 사용

- 마침표와 물음표를 사용하고, 느낌표나 쉼표는 사용하지 않는다.
- 마침표는 문장이 완전히 끝났을 때에만 사용하며, 문장을 끝맺지 못하였을 경우는 단위를 나눈 경우라도 마침표를 붙이지 않는다. 이때 문장이 완전히 끝났다는 것은 '-다', '-어요', '-어라' 등 종결어미로 끝났음을 뜻한다.

〈표 Ⅲ-22〉 방언 전사-표준어 대응쌍 전사 예시

올바른 표기	잘못된 표기
뉴스룸의 앵커브리핑을 시작하겠습니다.	뉴스룸의 앵커브리핑을 시작하겠습니다!
중화요리 집에서 홀서빙을 했었습니다.	중화요리 집에서, 홀서빙을 했었습니다.
저번 주에 그만두게 되었는데 이제	저번 주에 그만두게 되었는데 이제,

- 문장이 종결되었을 경우 마침표 또는 물음표를 반드시 붙이며, 마침표나 물음표를 붙인 경우는 반드시 단위를 나눈다.

〈표 Ⅲ-23〉 방언 전사-표준어 대응쌍 전사 예시

올바른 표기	잘못된 표기
똑같이 했을 겁니다. 어~ 그게 이거는 습관이고 버릇이고	똑같이 했을 겁니다. 어~ 그게 이거는 습관이고 버릇이고

- 말끝을 올리는 경우는 물음표를 붙인다. 특히 '-어', '-어요' 등 말끝을 올리거나 내리는 것에 따라 의미가 달라지는 경우, 반드시 마침표와 물음표를 사용하여 구분해 준다.

〈표 Ⅲ-24〉 방언 전사-표준어 대응쌍 전사 예시

평서문	의문문
-	그냥 월급 루팡이 되는 듯한 기분?
밥은 먹었어.	밥은 먹었어?

- 대화 순서 겹침
  - 대화의 순서가 겹쳐 동시에 소리가 들리는 경우는 따로 표시하지 않고 대화를 먼저 시작한 화자를 기준으로 시간 순서에 따라 적는다. 만약 상대방의 맞장구를 치는 표현(예: 네, 그렇죠, 맞아)이 중간에 나오면 앞선 대화를 완전히 적은 다음에 맞장구치는 표현을 줄을 바꾸어 적는다.

〈표 Ⅲ-25〉 대화 순서 겹침 전사 예시

구분	보기
주 발화	1: 딸 하나 (나)/(낳아) 갖고
맞장구 발화	2: 네.
주 발화	3: 세 살 (묵아)/(먹어) (잊아뿌고)/(잊어버리고)

- 끊어진 단어(단어가 불완전하게 발화된 경우)
  - 단어가 완전히 발음되지 않고 끊어진 경우는 그대로 전사하고 아래 예와 같이 –를 해당 단어 앞뒤에 붙인다. 이러한 단어가 둘 이상인 경우에는 모두 –를 붙인다.

〈표 Ⅲ-26〉 끊어진 단어 전사 예시①

올바른 표기	잘못된 표기
–전– –전– 전통이라고 우리가 흔히 얘기할 때	전통이라고 우리가 흔히 얘기할 때 전 전 전통이라고 우리가 흔히 얘기할 때
–학– 학교 아니 유치원에	–학– –학교– 아니 유치원에

  - 발화자가 실수로 원래 하려던 말이 아닌 다른 말을 한 뒤 올바르게 수정한 경우, 또는 같은 단어를 반복해서 말한 경우에는 -를 표시하지 않고 들리는 그대로 표기한다. 그리고 해당 음성 단위에는 반드시 코멘트를 달아 놓도록 한다.

〈표 Ⅲ-27〉 끊어진 단어 전사 예시②

올바른 표기	잘못된 표기
–학– 학교 아니 유치원에	–학– –학교– 아니 유치원에
크라운 –베– 크라운 베이커리 생크림이 좀 맛있죠.	–크라운– –베– –크라운– 베이커리 생크림이 좀 맛있죠.

- 띄어쓰기
  - 띄어쓰기는 띄어쓰기 규정에 맞게 한다
  - 의존명사는 띄어 쓰고, 수를 적을 때는 만 단위로 띄어 쓴다(예: 십이억 삼천백만 팔백구 불 등). 판단하기 어려운 경우에는 수시로 논의하여 결정한다(예: 오십대, 일 대 이)
  - 본용언과 보조용언은 띄어 쓴다(예: 먹어 버리다, 가고 싶다, 먹지 못하다)

〈표 Ⅲ-28〉 띄어쓰기 전사 예시①

올바른 표기	잘못된 표기
많이 먹는구나 그걸로 넘어지기야 하겠냐만은	많이 먹는∨구나 그걸로 넘어∨지기야 하겠냐만은

※ [주의] 단어를 발음하는 중간에 쉼이 들어간 경우에는 띄어 쓰지 않는다.

〈표 Ⅲ-29〉 띄어쓰기 전사 예시②

올바른 표기	잘못된 표기
뭐라카노	뭐라∨카노

※ [주의] 방언이 축약되어 띄어쓰기 규정을 적용하기 어려운 경우는 붙여 쓴다. 예를 들어 경상방언 '뭐라카노'의 '-카-'는 '(뭐라)고 하(노)'의 축약형이다.

- 축약형의 표기

  - 구어에서는 발음의 축약 현상이 많이 나타나는데, 두 음절이 한 음절이 될 수도 있고 네 음절이 두 음절로 줄어들 수도 있다. 이때 발음된 음절수와 표기상의 음절수를 맞추는 것을 원칙으로 한다. 따라서 축약형의 경우 모두 전사 표기에 반영한다. 표준어 대응쌍에서는 원래 형태를 다시 밝혀 적는다.

〈표 Ⅲ-30〉 축약형 전사 예시①

축약 대상	보기
그냥	(걍)/(그냥)
그러니까	(그니까)/(그러니까)

※ [주의] 〈표준국어대사전〉에 준말로 등재되어 있는 다음과 같은 단어들은 일일이 표준어 대응쌍을 적지 않는다.
예) 근데, 얘기, 요새, 애, 담, 맘, 첨, 낼, 젤, 좀, 재밌다, 갖다, ...

  - 사귀어, 바뀌어 등에서 모음 '위'와 '어'가 합쳐져 1음절로 축약되어 발음되는 경우는 다음과 같이 모음 'ㅕ'로 바꾸어 전사한다. 표준어 대응쌍에서는 원래 형태를 다시 밝혀 적는다.

〈표 Ⅲ-31〉 축약형 전사 예시②

축약 대상	보기
사귀어	(사겨)/(사귀어)
바뀌어	(바껴)/(바뀌어)
뛰어	(뗘)/(뛰어)

● 담화 표지

- “담화 표지”는 대화 상황에서 말하는 이가 머뭇거림, 이야기를 계속 하고 싶어 하는 등의 의도나 심리적 태도를 전달하기 위해 사용하는 것이다. 여기서는 “이, 그, 저, 아, 어, 에, 음, 응, 뭐” 등 1음절 담화표지에 한해서만 본래의 품사와 구별하기 위해 물결표(~)를 붙여 전사한다.
- 여기서 물결표(~)를 같이 전사하는 경우는 머뭇거림의 느낌을 주는 담화표지이다. 즉, “이 사람, 그 사람, 저 사람”처럼 가리키는 말로 쓰이는 “이, 그, 저”나 감탄의 “아, 어, 에, 음, 응, 뭐” 등이 원래의 의미로 쓰이지 않고, 말을 더듬거리거나 머뭇거릴 때 사용될 경우에만 물결표(~)를 붙여 표기한다. (“인제, 이제, 그냥, 무슨, 어떤” 등은 2음절이므로 물결표(~)를 붙이지 않음).

〈표 Ⅲ-32〉 담화 표지 전사 예시

지시·감탄의 경우	담화 표지인 경우
그 돈 벌고 싶어서	그~ 돈 벌고 싶어서
그냥 저 통상적인 노하우인지	그냥 저~ 통상적인 노하우인지
응. 얼마만인지 모르겠네	응~ 얼마만인지 모르겠네

● 잘 들리지 않는 부분

- 대화가 잘 들리지 않는 부분이나 잘 들리지 않지만 해당 부분을 추측할 수 있는 경우, 추측한 내용을 (()) 안에 적는다.
- 잘 들리지 않는 부분 중 일부분만 들리거나 추측 가능한 경우, 추측 가능한 부분을 (()) 안에 적되, 들리지 않는 부분은 그 음절 수만큼 x로 나타낸다.

〈표 Ⅲ-33〉 잘 들리지 않는 부분 전사 예시

구분	보기
추정 불가능	**(())** 너무한 거 같더라.
추정 가능	그 전까지는 직장 생활 (하니라구)/(하느라고) **((더 힘들어))**
일부 추정 가능	그거 진짜 **((xx해야))** 되겠더라.

- 잘 들리지 않는 부분을 전사하기 위해 반복 청취 등의 노력을 들이지 말고 가급적 (()) 처리하는 것이 바람직하다.

- 말소리를 제외한 기타 소리들
  - 웃음, 목청 가다듬는 소리, 박수, 노래 등 직접적인 말소리가 아닌 소리에는 @를 앞에 붙여서 다음과 같이 적는다. 전사 후 최종 단계에서는 다음과 같이 마크업된다.

〈표 Ⅲ-34〉 기타소리(말소리 외) 전사 예시

구분	전사	마크업
웃음	@웃음	{laughing}
목청 가다듬는 소리	@목청	{clearing}
박수	@박수	{applauding}
노래	@노래	{singing}

  - 위 4가지 외에 기침, 들숨, 날숨, 재채기, 코훌쩍임, 하품 등의 소리는 전사하지 않는다.

- 개인 정보의 보호
  - 대화자들의 신분 보장을 위해 이름, 주민등록번호, 카드 번호, 전화번호 등 개인정보와 관련된 사항은 노출되지 않도록 전사 단계에서 표시를 해 둔다.

〈표 Ⅲ-35〉 개인정보 전사 예시

구분	전사	마크업
이름	@이름	&name&
상호명	@상호명	&company-name&
주민등록번호	@주민번호	&social-security-num&
카드번호	@카드번호	&card-num&
주소	@주소	&address&
전화번호	@전화번호	&tel-num&

※ [주의] 정치인, 연예인 등 유명인의 이름은 위와 같이 하지 않고 그대로 전사한다.
※ [주의] 공적 성격을 지닌 이름들의 경우도 비식별화하지 않고 그대로 전사한다. 예를 들어 학교, 기관명, 단체명, 영화 제목, 노래 제목, 책 제목, 방송 제목, 게임명, 상품명, 제품명 등이 있다.

● 특정 상호가 발화되었을 경우 그대로 전사하지 않고 아래와 같이 적는다.

신촌에 @상호명은 진짜 맛없어.

※ [주의] 넷플릭스, 유튜브, 삼성, 엘지, 애플 등 널리 알려져 있는 상호에 대해서는 위와 같이 하지 않고 그대로 전사한다. 개인상호만 위와 같이 표시한다.
※ [주의] 개인 유튜브 채널명도 신분 보장을 위해 비식별화해야 하며, 이때 이름이 아닌 상호로 취급하여 '@상호명'으로 표시한다.

뭐~ @상호명1나 아니면 @상호명2 이런 거 자주 봐.

- 주소는 동 이하의 구체적인 주소만 표시하며, 동 이상의 주소는 그대로 전사한다.

근데 너 연희동 살잖아.(o)

- 여러 이름이 나올 때는 번호를 붙여 구별해야 한다. 이때 한 파일 내에서 해당 번호가 가리키는 대상이 일관성을 지녀야 한다.

〈표 Ⅲ-36〉 특정 상호 전사 예시

그때 철수랑 민수랑 너랑 나랑 갔잖아. 철수도 알고 있지?

올바른 표기	잘못된 표기
그때 @이름1이랑 @이름2이랑 너랑 나랑 갔잖아. @이름1도 알고 있지?	그때 @이름1이랑 @이름2이랑 너랑 나랑 갔잖아. @이름3도 알고 있지?

※ [주의] 담화 안에서 등장하는 사람 이름이 매우 많을 수 있으므로, 전사 작업 시 반드시 이름과 순번의 대응을 따로 기록해 두고 헷갈리지 않도록 주의한다.
※ [주의] 상호명의 경우도 여러 상호명이 나오면 상호명1, 상호명2, ...로 구별하여 전사한다.

- 받침이 없는 이름 뒤에 '-입니다', '-인데' 등이 결합해서 '-ㅂ니다', '-ㄴ데' 형태로 나타나더라도 '-입니다', '-인데' 등을 살려 쓴다.

〈표 Ⅲ-37〉 이름 뒤 받침 없는 경우 전사 예시

저는 김철숩니다.
나 김민순데 알지?

올바른 표기	잘못된 표기
저는 @이름1입니다.	저는 @이름1ㅂ니다.
나 @이름2인데 알지?	나 @이름2ㄴ데 알지?

- 기타 지침
  - 방언 전사를 위해 사용한 기호(예: -, {}, &, (()) 등)는 표준어 대응쌍 표기에는 사용하지 않는다.
  - 큰따옴표나 작은따옴표를 사용하지 않는다. 즉 발화자가 글을 읽을 때 따옴표로 표시하지 않고 내용 전사만 한다.
  - 말꼬리를 끌어 장음으로 발음한 경우에 이를 반영하지 않고 원래의 단어로 적는다.
    예) 학스읍(x) => 학습, 소오름(x) => 소름
- 지역별 방언 전사 주의 사항

1) 경상방언

- 종결어미에 '-이'가 결합한 '-대이, -래이, -재이'은 소리대로 적는다.

집에 (갔대이.)/(갔다.) 전화 (해래이.)/(해라.) 다음에 (보재이.)/(보자.)

- 표준어의 '그러다'에 해당하는 '그카다, 그쿠다' 등은 소리대로 적는다.

(그카면)/(그러면) 저기 갔다 올 (끼가?)/(거가?) (그쿠면)/(그러면) 그 일은 (끝났나?)/(끝났니?) (그카고)/(그러고) 있지 말고 (일로)/(이리로) (온나)/(오너라) (그쿠고)/(그러고) 잘난척은 (자가)/(쟤가) 잘한다.

- 받침 'ㅇ'이나 'ㄴ'이 나타나지 않으면 소리대로 적는다.

(주머이)/(주머니) (어무이)/(어머니) (사이)/(산이) 크다 (학새이)/(학생이)

2) 전라방언

- 표준어 '-으니까'의 방언형은 '-응께, -응게, -응께네, -으니께' 등으로 소리대로 적는다.

(인자는)/(인제는) 약이 (조응께)/(좋으니까) 그 공식적으로 다 (허니께)/(하니까)

- 둘째 음절 이하의 'ㅎ'이 나타나지 않는 말의 경우, 다음과 같이 소리대로 적는다.

(뭐다러)/(뭣 하러) 그러냐?
잘 하지도 (모다고)/(못하고)
(배과점에)/(백화점에) 가 (봉께)/(보니까)
벌써 (유강년이여?)/(육 학년이야?)
으메 (답다번거)/(답답한 것)
눈 앞이 (깝까버다.)/(갑갑하다.)

- 'ㄴ'이 나타나지 않으면 소리대로 적는다.

(가마이)/(가만히)
(마이씩)/(많이씩)
(아잉)/(아닌) 게 (아이라)/(아니라)

3) 제주방언

- 앞에 요소가 받침이 있는 음절이고 후행하는 요소가 모음으로 시작할 때 후행 단어의 첫 음절 자리에 받침 자음을 복사하여 발음한다. 이러한 경우는 소리대로 적는다.

(한국금식)/(한국음식)
(맏다덜)/(맏아들)
(비단놋)/(비단옷)
(감못)/(감옷)

- ᄋᆞ(아래아) 는 '아*'로 적는다.

(밟*아)/밟아
(나*말*)/(나물)
(아달*)/(아들)

※ [참고] 고동호(2008: 69-70),「제주방언 ·의 세대별 변화 양상」,『한국언어문학』65, 55-74., 김원보(2006: 134-135),「제주방언화자의 세대별(20대, 50대, 70대) 단모음의 음향분석과 모음체계」,『언어과학연구』39, 125-136. 등 최근 연구에서는 제주방언에서 70대 이상은 [ᄋᆞ] 발음을 유지하고 있고, 50대는 개인에 따라 어휘에 따라 있기도 없기도 하고, 20대는 [ᄋᆞ] 발음이 거의 없다고 보고하였다.

4) 충청방언

- 종결어미 '-댜'는 소리대로 적는다.

그 낮에 꿈을 (꾸니께)/(꾸니까) (그라댜)/그러더래)
우리 (야덜)/(아들) (잡어)/(잡아) (간댜.)/(간대.)

- 표준어 '어떻게'의 방언형은 '워떻게, 어티기' 등은 소리대로 적는다.

> 혹시 (워떻게)/(어떻게) 하는 건 줄 아세요?
> 장사 하려고 (어티기)/(어떻게) 집을 크게 (졌는디)/(졌는데)

5) 기타

- (어두 된소리화 현상) 방언에서 흔히 나타나는 어두 된소리화의 경우, 방언의 특성으로 볼 수 있으므로 소리나는 대로 전사하고, 표준어 대응쌍 이중전사를 한다.

> 예) (쩌번에)/(저번에), (따르다)/(다르다), (꼐속)/(계속)

● 표준어 대응쌍 전사지침

- 기본 원칙

1) 현재 컴퓨터를 사용하여 한국어를 분석하는 도구는 기본적으로 표준어를 기반으로 개발되었다. 방언과 함께 표준어 대응쌍을 구축하는 것은 특수한 방언형과 표준어를 쌍으로 제시함으로써 표준어의 규칙을 벗어나는 구어 방언 데이터를 AI가 인식할 수 있도록 하기 위함이다.

2) '방언 전사'와 '표준어 대응쌍'은 "(방언 전사 형태)/(표준어 대응쌍 형태)"의 꼴로 제시한다. "방언 전사"를 먼저하고 "표준어 대응쌍 전사"를 그 뒤에 나란히 제시한다.

〈표 Ⅲ-38〉 방언-표준어 대응쌍 전사 예시

지역	보기
강원	어제는 막 아프다고 (그라)/(그렇게) 난리를 치더니 오늘은 좀 괜찮냐?
경상	고등학교 (댕길)/(다닐) 때 미역 (쭐거리)/(줄기) 반찬도 (마이)/(많이) (묵었지.)/(먹었지.)
전라	나는 (거까정은)/(거기까지는) 잘 (모릉께)/(모르니까) 이제 더 묻지 마시오.
제주	보리 (한*)/(한) 말이면 부인들 (비렁)/(빌려) (했주게.)/(했지).
충청	그럼 내가 (지비)/(집에) 갔다 올 (때꺼정)/(때까지) (지다리실텨?)/(기다리실 테요?)

3) 표준어 대응쌍의 작성은 국어 어문규범(한글 맞춤법, 표준어 규정, 외래어 표기법, 로마자 표기법)에 따른다.

4) 문장부호는 괄호 속에 넣어서 방언 전사형과 표준어 대응쌍 모두에 동일하게 제시한다.

〈표 Ⅲ-39〉 문장부호 포함한 방언-표준어 대응쌍 전사 예시

지역	보기
강원	나도 이쪽 동네 (출신이래.)/(출신이야.) 니가 말한 건 이게 (다나?)/(다니?)
경상	떡을 (맹갈아)/(만들어) (묵었지.)/(먹었지.) 어제 어디 (갔었노?)/(갔었니?)
전라	혼자 다 (묵어 분당께.)/(먹어 버린다니까.) 난장을 쳐도 (쪼깨)/(조금) (늦겄는디?)/(늦겠는데?)
제주	하루에 같이 (검질멧주게.)/(김매었지). (게민)/(그러면) (멋을)/(뭣을) 더 (가*라 주코?)/(말해 줄까?)
충청	(오둥아를)/(오디를) 먹었는지 입 안이 시커멓게 (물들었슈.)/(물들었어요.) 동네 사람들은 (워떡헌댜?)/(어떡한대?)

● 표준어 대응쌍 표기와 선별 기준

1) 방언형에 대응되는 표준어 대응쌍 표기는 국립국어원의 〈우리말샘〉의 정보를 기준으로 삼는다. 아래에서 방언 '맹갈다, 묵다'에 대해 '만들다, 먹다'를 표준어 대응쌍으로 전사한다.

※ 〈우리말샘〉 https://opendic.korean.go.kr/main

〈표 Ⅲ-40〉 우리말샘 기준 방언-표준어 대응쌍 전사 예시

지역	보기
강원	돈이 (읎어도)/(없어도) 남한테 (아수운)/(아쉬운) 소리는 못하겠다. 〈우리말샘〉 읎다 아숩다 • 읎다 「001」「형용사」「방언」 '없다'의 방언(강원, 경기, 전북, 충청) • 아숩다 「001」「형용사」「방언」 '아쉽다'의 방언(강원, 경상, 전라, 충청)
경상	떡을 (맹갈아)/(만들어) (묵었지.)/(먹었지.) 〈우리말샘〉 맹갈다 묵다 • 맹갈다 「001」「동사」「방언」 '만들다'의 방언(경남, 전남) • 묵다 「004」「동사」「방언」 '먹다'의 방언(강원, 경상, 전라)
전라	그거 다 (이애기헐라면)/(이야기하려면) (미칠을)/(며칠을) 해도 안 돼. 〈우리말샘〉 이애기 미칠 • 이애기 「001」「명사」「방언」 '이야기'의 방언(강원, 경북, 전라) • 미칠 「001」「명사」「방언」 '며칠'의 방언(경남, 전남)
제주	성격이 (잘도)/(매우) (요망지다.)/(야무지다.) 〈우리말샘〉 잘도 요망-지다 • 잘도 「001」「부사」「방언」 '매우'의 방언(제주) • 요망-지다 「001」「형용사」「방언」 '야무지다'의 방언(제주)

지역	보기
충청	그럼 내가 갔다 올 (때꺼정)/(때까지) (지다리실텨?)/(기다리실 테요?) 〈우리말샘〉 꺼정 지다리다 * 꺼정 「001」「조사」「방언」 '까지'의 방언(강원, 경상, 충청, 함경) * 지다리다 「001」「동사」「방언」 '기다리다'의 방언(충청)

※ 대화 상황에서 자주 사용되는 중앙방언형 어미 '-애', '-어', '-두', '-구' 등은 비표준이지만 예외적으로 표준어 대응쌍을 제시하지 않는다, 다만 어간에 방언형이 나타나 표준어 대응쌍을 제시할 경우에는 어미 '-애', '-어', '-두', '-구' 등도 함께 표준어를 제시한다.

〈표 Ⅲ-41〉 중앙방언형 어미 전사 예시

올바른 표기	잘못된 표기
진짜 우리가 잘 되길 바래.	진짜 우리가 잘 되길 (바래.)/(바라.)
너무 많이 지난 것 같애.	너무 많이 지난 것 (같애.)/(같아.)
그런 생각은 아예 하지를 말어.	그런 생각은 아예 하지를 (말어.)/(말아.)
그 사람 말이 맞어.	그 사람 말이 (맞어.)/(맞아.)
이제 나는 하나두 모르겠다.	이제 나는 (하나두)/(하나도) 모르겠다.
손부터 씻구 밥을 먹어라.	손부터 (씻구)/(씻고) 밥을 먹어라.
아직도 안 돌아갔다구?	아직도 안 (돌아갔다구?)/(돌아갔다고?)
이제 밥을 먹으려구 한다.	이제 밥을 (먹으려구)/(먹으려고) 한다.
이래 (가주구)/(가지고)	이래 (가주구)/(가지구)

2) 여러 단어가 합쳐져서 만들어진 말의 경우 사전 표제어 형식을 확인한다. 두 단어가 -로 연결된 경우 전체를 한 덩어리로 간주한다. 두 단어가 ^로 연결된 경우나 연결되지 않은 경우는 각각의 단어에 대해 표준어 대응쌍을 표기한다.

〈표 Ⅲ-42〉 여러 단어가 합쳐진 경우 전사 예시

천지-삐까리 (天地삐까리) * 천지-삐까리 「001」「명사」「방언」 매우 많음(경상).
서답 구덕 * 서답 구덕 「001」「방언」 '빨래 바구니'의 방언(제주).

올바른 표기	잘못된 표기
할 게 (천지삐까리다.)/(매우 많다.)	할 게 (천지)/(매우) (삐까리다.)/(많다.)
(서답)/(빨래) (구덕에)/(바구니에) 담아	(서답 구덕에)/(빨래 바구니에) 담아

3) 방언형이 여러 개의 표준어와 대응되는 경우에는 다음과 같은 원칙에 따른다.

- 형태적으로 가장 가까운 표준형을 선택한다.

〈표 Ⅲ-43〉 방언형-표준형 전사 예시

올바른 표기	잘못된 표기
지금 (머라카노?)/(뭐라고 하니?)	지금 (머라카노?)/(뭐라는 거니?)

- 대화 상황에서 어미는 대화 맥락이나 어감, 화자·청자의 관계 등에 따라 어울리는 표준형이 달라질 수 있다. 이러한 경우 상황에 맞게 표준어형을 선택하며, 이 부분에 대한 판단은 전사자에 따라 다소 주관적일 수 있다.

〈표 Ⅲ-44〉 상황에 맞는 표준형 전사 예시

방언형	복수 대응 표준어형	사용 맥락
경상방언 '-노?'	어제 어디 (갔었노?)/(갔었냐?)	친구 등 동등한 관계에서 사용
	어제 어디 (갔었노?)/(갔었니?)	부자, 사제 등 상하 관계에서 사용
경상방언 '-꾸마'	내가 (하꾸마.)/(할게.)	친구 등 동등한 관계에서 사용
	내가 (하꾸마.)/(하마.)	부자, 사제 등 상하 관계에서 사용
충청방언 '기여(겨)', '그려'	그게 (기여?)/(맞아?) 그게 (기야?)/(맞아?) 그게 (기냐?)/(맞냐?)	내용을 확인하는 질문에서 사용
	(겨)/(그래) (아녀?)/(안 그래?) (기여)/(그래) (아니여?)/(안 그래?)	'겨 아녀?'의 맥락에서 사용
	(그려.)/(그래.) (겨.)/(그래.) (아녀.)/(아니야.)	질문에 대답하는 상황에서 사용

※ 이것은 맥락에 따라 어감이 달라지는 어미에만 해당하며, 명사나 동사 등 단어의 경우에는 맥락상 다소 어색하더라도 사전에 제시된 형태를 그대로 따른다.

〈표 Ⅲ-45〉 어감이 다른 경우 표준형 전사 예시

머스마
• 머스마 「001」「명사」「방언」 '사내아이'의 방언(강원, 경상, 전북, 충청).

올바른 표기	잘못된 표기
그 직원이 (머스만데.)/(사내아인데.)	그 직원이 (머스만데.)/(남잔데.)

4) 방언형에 형태적으로 유사한 표준형이 없을 때는 의미적으로 가장 유사한 표준형을 선택한다. 예를 들어 경상방언의 '-매로', '-맨치로'는 형태적으로 비슷한 표준어 어휘가 없지만, 의미적으로는 '-처럼'과 거의 동일하므로 '-처럼'을 표준어 대응형으로 적는다.

(니매로)/(너처럼)
(니맨치로)/(너처럼)

5) 형태적으로도 의미적으로도 대응 표준어를 사전에서 발견할 수 없는 경우, 대응하는 표준형이 없는 것으로 간주한다. 이러한 경우 표준형을 입력할 자리에 '#방언형'을 입력하며, 방언형의 뜻을 풀이하여 표준형에 기술하지는 않는다.

〈표 Ⅲ-46〉 방언형이 표준어 사전에서 발견할 수 없는 경우감이 다른 경우

올바른 표기	잘못된 표기
(뭉티기)/(#뭉티기) (무러)/(먹으러) 가자	(뭉티기)/(생고기를 엄지손가락만하게 썰어 낸 음식) (무러)/(먹으러) 가자
(하고재비가)/(#하고재비가)	(하고재비가)/(하고 싶어 하는 사람이)

6) 방언형에 대한 표준어 대응쌍은 가급적 음절 및 어절 수를 맞추어 제시한다. 그러나 음절 및 어절 수를 맞출 수 없는 경우도 있는데, 예로 아래 '경상, 충청'처럼 한 어절이 두 어절의 표준어 대응쌍을 가질 수도 있다.

〈표 Ⅲ-47〉 음절 및 어절 수 다른 경우 표준형 전사 예시

지역	보기
강원	어제는 막 아프다고 (그라)/(그렇게) 난리를 치더니 오늘은 좀 괜찮냐?
경상	지금 (머라카노?)/(뭐라고 하니?)
전라	혼자 다 (묵어)/(먹어) (분당께.)/(버린다니까.)
제주	하루에 같이 (검질멧주게.)/(김매었지.)
충청	언제까지 (지다리실텨?)/(기다리실 테요?)

- 축약과 생략

1) 방언전사에서 탈락된 소리는 표준어 대응쌍에서 복원시켜 본딧말로 바꾸어 적는다.

전부 (수매르)/(수매를)
(하꺼인디)/(할 것인데)

2) 방언에서 축약형으로 표기된 형태 역시 표준어 대응쌍에서 본딧말로 바꾸어 적는다. 축약형와 본딧말 모두 표준어로 사전에 등재되고, 문맥상 축약형이 본딧말보다 더 자연스러울 경우, 축약형도 허용한다. 예를 들어 '거'와 '것'은 모두 표준형 대응쌍에 사용할 수 있지만, '할 거인데(×)'와 같이 문맥 상 '거'로 바꾸는 것이 허용되지 않는 경우는 반드시 '것'으로 적어야 한다.

〈표 Ⅲ-48〉 축약 및 생략이 있는 방언 전사 예시

올바른 표기	잘못된 표기
(하꺼인디)/(할 것인데)	(하꺼인디)/(할 거인데)

3) 〈우리말샘〉에 준말로 등재되어 있는 다음과 같은 단어들은 일일이 표준어 대응쌍을 적지 않는다.

예) 근데, 얘기, 요새, 애, 담, 맘, 첨, 낼, 젤, 좀, 재밌다, 갖다, …

4) 외래어, 외국어의 준말의 경우 다른 외래어, 외국어와 마찬가지로 표준어 대응쌍을 적지 않는다. 준말의 발화 그대로 전사한다.

예) 알바, 페북, …

- 띄어쓰기
  - 방언형에서는 띄어쓰기가 무시되었더라도 표준형에서는 어문규범에 준하여 띄어쓴다.

(뭐라카노?)/(뭐라고 하니?)
(뭐라캐)/(뭐라고 해) (쌓노)/(쌓니)
(이래)/(이렇게 해) (가주고)/(가지고) 농사를 (지야)/(지어) (노으면)/(놓으면)

- 방언권별 조사, 어미 표준어 대응쌍 목록(일부)
  - 〈우리말샘〉의 방언(조사, 어미, 품사없음)에 대한 표준어 대응쌍의 일부를 보이면 다음과 같다. 목록 전체는 별첨 자료(우리말샘_방언_조사 어미 품사없음_5개권역)를 참고하면 된다.

〈표 Ⅲ-49〉 방어권별 방언-표준어 대응쌍

방언권	어휘	품사	표준어 대응쌍
강원	가	품사 없음	걔
	갠데	품사 없음	그런데
	-게르	어미	-게

방언권	어휘	품사	표준어 대응쌍
	아무케	품사 없음	아무렇게
	-았댔-	어미	-았었-
	야	품사 없음	얘
	오러	품사 없음	요렇게
	우뜨케	품사 없음	어떻게
	우째	품사 없음	어째
	울-만큼	품사 없음	얼마큼
	이러	품사 없음	이렇게
	인데	조사	한테
	자	품사 없음	쟤
	처름	조사	처럼
전라	-간디	어미	-관데
	거따가	품사 없음	게다가
	거러치럼	품사 없음	그렇게
	고러치롬	품사 없음	그렇게
	-그레	어미	-기에
	까장	조사	까지
	-당께로	어미	-다니까
	땜시	품사 없음	때문에
	-라우	어미	-요
	매이로	조사	처럼
	-ㅂ디껴	어미	-ㅂ디까
	아무케	품사 없음	아무렇게
충청	-것-	어미	-겠-
	겨	품사 없음	거야
	그랴도	품사 없음	그래도
	까장	조사	까지
	-ㄹ라구	어미	-려고
	-ㄹ텨	어미	-ㄹ래
	-어유	어미	-아요
	워떡-허다	품사 없음	어떡하다
	워째서	품사 없음	어찌하여서

방언권	어휘	품사	표준어 대응쌍
	워척-허다	품사 없음	어떡하다
	워치게	품사 없음	어떻게
	이릏게	품사 없음	이렇게
	튜	품사 없음	테요
제주	-게겐	어미	-자꾸나
	게고-제고	품사 없음	그러고저러고
	드레	조사	으로
	-멍	어미	-면서
	-메서란	어미	-매
	-센	품사 없음	-라고
	신디	조사	한테
	아메-나	품사 없음	아무렇게나
	-앙근에	어미	-고서, -아서
	야이	품사 없음	얘
	-어근	어미	-어서
경상	까짐	조사	까지
	-꺼마	어미	-ㄹ게
	-꾸마	어미	-마
	끈	조사	까지
	-는기라	품사 없음	-는 거야
	-니껴	어미	-ㅂ니까
	-래이	어미	-라
	매추로	조사	처럼
	-시다	어미	-오
	-시소	어미	-십시오
	우짜모	품사 없음	어쩌면
	-응께	어미	-으니까
	이카면	품사 없음	이렇게 하면

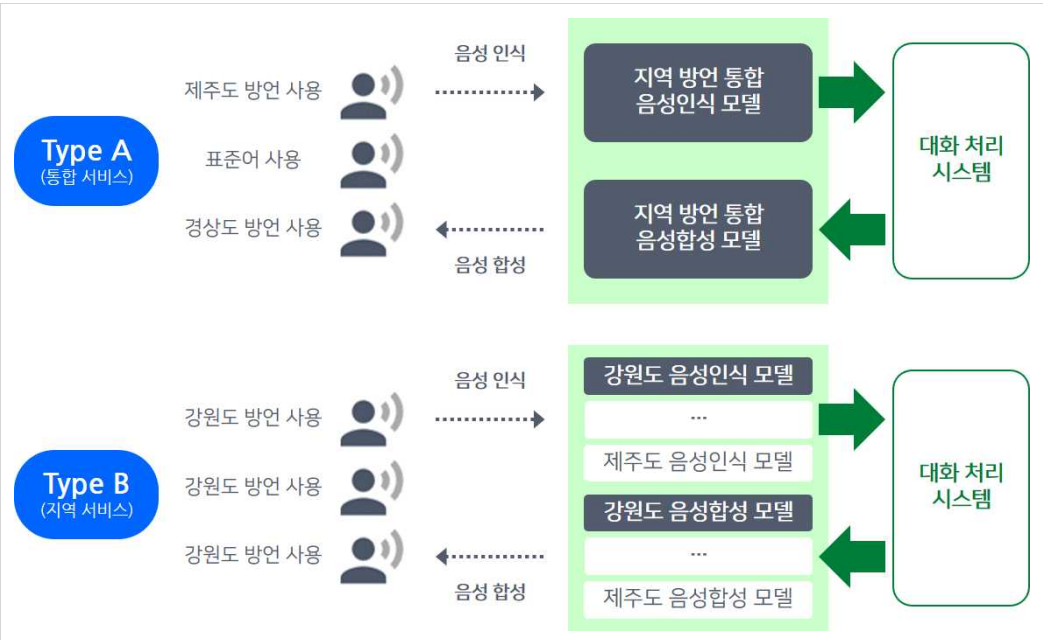
5 데이터 활용 방안

5.1 학습 모델

〈표 Ⅲ-50〉 한국어 방언 발화 데이터 학습 모델 및 성능 지표

과명제	AI 모델	모델 성능 지표
한국어 방언 발화 데이터 (강원도) 한국어 방언 발화 데이터 (경상도) 한국어 방언 발화 데이터 (전라도) 한국어 방언 발화 데이터 (제주도) 한국어 방언 발화 데이터 (충청도)	• 각 도별 음성인식 모델 (각 도별 5개, 5도 통합 1개) • 각 도별 음성합성 모델 (5개) • 각 도별 방언-표준어 기계번역 모델 (각 도별 5개, 5도 통합 1개) • GPT 기반 자연어 생성 기술을 이용한 일상 대화 모델 (각 도별 5개, 5도 통합 1개)	• 음성인식 모델 음절 인식율1) • 음성합성 모델 음성 품질 수치 (MOS2): Mean opinion score) • 기계번역 모델 번역 품질 수치 (BLEU3): bilingual evaluation understudy) • 일상 대화 모델 대화 품질 수치 (Perplexity4); SSA5): Sensibleness and Specify Average, 구글에서 제안한 Human Evaluation Metric)

5.2 서비스 활용 시나리오



[그림 Ⅲ-55] 한국어 방언 발화 데이터 활용 시나리오

제5장 한국어 SNS 데이터

1 데이터 정보 요약

1.1 가이드 분류

대분류	텍스트	중분류	자연어	소분류	SNS

1.2 데이터 정보

데이터 이름	한국어 SNS 데이터
데이터 요약	SNS 상에서 이루어진 한국어 대화 데이터를 수집하여 정제 및 개인정보 비식별화 작업을 거친 후 화자 정보 등의 메타 데이터를 부가한 총 200만 건의 학습 데이터를 제공함
데이터 출처	크라우드 소싱에 의한 수집한 카카오톡 대화 데이터

1.3 데이터 구축 개요

- 데이터 구축 목표: 한국어 구어체 텍스트 기반의 정보검색, 대화분석, 질의응답, 명령어 이해, 언어모델 학습 등의 자연어처리 AI 기술 개발을 위한 한국인의 일상대화 메신저 채팅 데이터 구축
- 데이터 구축 내용: 범용 모바일 메신저인 카카오톡 메신저 대화 원문 수집
- 데이터 규모: 적정 길이의 정제된 입력 텍스트 대화세트 200만 건

● 데이터 구축 절차

〈표 Ⅲ-51〉 한국어 SNS 데이터 구축 절차

단계	세부 절차	설명	산출물
준비	작업 환경 구축	작업 도구 선정	
	작업 대상 선정	획득할 데이터의 규격 및 조건 선정	
	데이터 제공 기관 검토	작업자 모집 기관 검토	

단계	세부 절차	설명	산출물
	작업자 확정	원시 데이터 작업자 및 제공자와 계약 체결	개인정보수집 및 이용 동의서, 근로계약서, 저작권활용계약서
	작업 지침서 작성	작업 지침서 및 가이드 동영상 제작	작업 지침서 가이드 동영상
획득	원시 데이터 획득	카카오톡 대화문(텍스트) 형태의 원시 데이터 획득	원시 데이터
정제	부적합 데이터 선별	데이터 수집 요건 미충족 대화 제외 중복 데이터 제외	요건 미충족 및 중복 제외 데이터
	데이터 비식별화	개인정보 마스킹 및 비식별화 민감정보 등의 삭제	비식별화 데이터
가공	작업 인력 교육	데이터 가공 작업 교육	
	유형 및 주제 구분	데이터의 유형 및 주제 구분 작업	가공 데이터
검사	SNS 데이터 검수	원문 및 라벨링 검수	검수 완료 데이터
	대화문 데이터셋 구성		대화 데이터셋
	외부 기관 품질 인증	관련 외부 기관의 품질 인증	품질 인증서
활용	AI 모델링 서비스		AI 모델링 소스
	데이터 개방	AI 허브 공개	AI 학습용 데이터

1.4 구축 목적

- 본 과제의 최종 임무는 범용 메신저를 이용하여 생성된 대화 원문에서 적정 기준과 조건을 만족하는 대화를 선별하여 정제한 뒤 정보검색, 대화분석, 질의응답, 명령어 이해, 언어모델 학습 등 자연어처리 AI 기술 개발을 위한 대규모 한국어 대화 텍스트 AI 데이터셋을 구축하는 것임

1.5 활용 분야

- 연구 분야: 한국인들이 일상생활 속 메신저를 통한 텍스트 커뮤니케이션에서 사용하는 대화 방식과 표현 및 어휘를 처리할 수 있는 언어모델 연구
- 정보검색, 대화 엔진, 질의응답, 명령어 이해 등 산업 분야: AI 상담센터, 챗봇, AI 스피커, 개인비서, 스마트홈 등 한국어 구어 자연어 처리 엔진이 필요한 산업

1.6 유의 사항

- 지적재산권 문제 해결
  - 본 데이터의 제한 없는 배포와 활용을 위해서는 대화 참여자들의 '개인정보 수집 및 이용 동의' 및 '저작권 활용 계약' 체결이 필수임
  - 과제 초기에는 구글폼(Google Forms) 등의 온라인 설문 도구를 이용하여 화자 정보 수집을 수행하고, 개인정보 수집 및 이용 동의와 저작권 활용 계약 체결은 온라인 전자 서명 플랫폼을 이용하여 진행하며, 이후에는 데이터 획득 및 저작 도구를 통하여 위의 과정을 진행함
  - 저작권 활용 계약은 저작물에 해당하는 대화 건별로 체결하는 것이 불가능하므로 대화 참여자가 본 과제에 제공하는 모든 대화에 대하여 저작권 이용을 허락하는 것으로 함
  - 제공된 대화 데이터는 과제 이후에 공개 및 활용되어야 하므로 저작권 활용 계약은 일정 시점(예: 2040년)까지 일정 주기(예: 5년)로 자동 갱신되도록 함
  - 제공된 대화 데이터는 연구와 기술 개발을 위해 공개되며, 복제, 변형, 분석이 가능함을 고지해야 함
  - 저작권 활용 계약 체결 당사자는 대화 데이터를 제공하는 대화 참여자와 본 사업의 참여 기관인 주식회사 바이브컴퍼니로 함
  - 다음 관련 법규에 대한 검토를 전문 기관에 의해 수행하여 법률적 문제 발생에 대처함

    ※ 저작권법, 저작권법 시행령, 저작권법 시행규칙

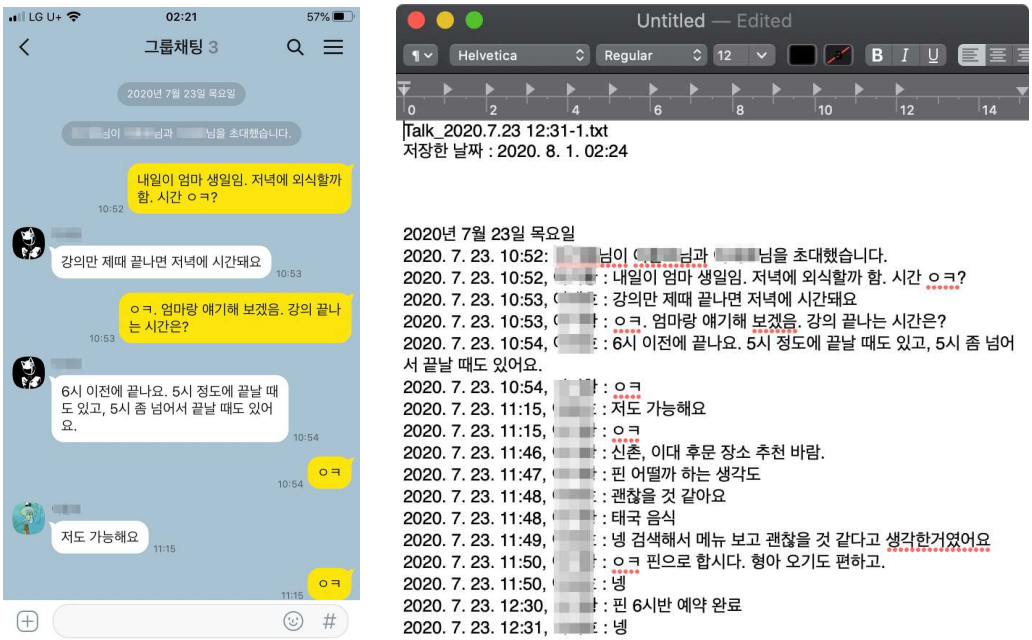
    ※ 개인정보 보호법, 개인정보 보호법 시행령, 개인정보 보호법 시행규칙
- 원시 데이터의 양품/불량 기준
  - 모든 대화 참여자의 개인정보 이용 동의와 저작권 활용 허락 계약 체결이 이루어지지 않은 대화 데이터는 데이터셋에 포함될 수 없음
  - 모든 대화 참여자의 화자 정보 제공이 이루어지지 않은 대화 데이터는 데이터셋에 포함하지 않음
  - 대화의 내용이 반사회적, 혐오적, 차별적, 선정적인 경우 데이터셋에 포함하지 않음
- 데이터 마스킹 및 개인정보 비식별화
  - 원시 데이터에 포함된 대화 참여자의 이름은 모두 익명화함

- 대화 내용에 포함된 기타 인명 등의 개인정보는 박일섭(2019)에서 제시한 비식별화 지침을 참고하여 마스킹 비식별화를 수행하되 해당 위치에 어떤 내용이 있었는지를 라벨링함
- 이름 범주에서 실명을 비롯하여 화자의 특정이 가능한 실명의 변형, 대화명, 필명 등은 모두 비식별화 대상이며 일반적인 애칭이나 연예인, 공인 등의 이름은 비식별화 대상이 아님
- 온라인 범주에 속한 아이디, 전자우편 주소와 각종 번호 및 비밀번호 범주에 속한 주민등록번호, 운전면허번호, 전화번호, 통장계좌번호 등의 모든 번호는 비식별화 대상이며, 지하철역명, 상호 등은 비식별화 대상이 아님
- 출신 및 소속 범주에 속하는 학교, 직장 등의 정보는 모두 비식별화 대상임
- 제시된 기준에 포함되지 않더라도 다른 정보와 결합하여 화자의 특정이 가능하다고 판단되는 정보들은 모두 비식별화함
- '씨*', '존*'와 같이 사용 빈도가 높고 대표적인 유형의 욕설이나 비속어에 대해 우선적으로 마스킹 처리를 수행하고, 그 이외에 욕설이나 비속어의 마스킹도 순차적으로 진행함

2 데이터 획득 및 정제

2.1 원시 데이터 선정

- 원시 데이터 형태
  - 원시 데이터는 카카오톡에서 이루어진 대화를 "대화 내용 내보내기" - "텍스트 메시지만 보내기" 기능에 의해 생성된 플레인 텍스트 파일 형태임



[그림 Ⅲ-56] 카카오톡 대화 내보내기의 실행 예시

- 원시 데이터 항목
  - 대화 참여자(화자)
  - 발화 수
  - 말차례 수
  - 화자 정보: 성별, 연령대, 거주지
  - 대화 원문
  - 발화 발생 날짜 및 시간

2.2 규제관련 사항

- 관련 내용 없음

※ 본 데이터 가이드라인 상에는 규제 관련 내용이 없지만, 데이터 구축 시 규제 관련 내용 확인 및 검토 필요

2.3 획득 및 정제 절차

1) 원시 데이터 획득 방안

- 데이터 획득 협력 기관의 회원들을 대상으로 크라우드 소싱에 의해 크라우드 워커를 모집
- 크라우드 워커 모집 단계에서 개인정보 수집 및 이용 동의와 저작권 활용 계약을 체결함
- 대화 데이터 수집 지침에 따라 대화 데이터의 기준, 대화 진행 방법, 데이터 제출 방법 등의 교육을 진행
- 크라우드 워커는 대화 주제 분류와 함께 추출된 텍스트 형식의 대화 데이터를 데이터 담당자에게 제출
- 원시 데이터의 수집은 구글폼(Google Forms)을 활용함
- 크라우드 워커가 구글폼으로 작성된 대화문 제공 신청서에서 이름, 연락처, 전자우편 주소, 계좌번호, 발화자 정보를 입력하여 제출함
- 신청서 작성을 완료한 크라우드 워커에게 개인정보 수집 및 이용 동의 계약서, 근로 계약서 및 저작권 활용 계약서를 온라인 서명 플랫폼을 통해 발송함
- 계약서 서명을 완료한 크라우드 워커는 카카오톡 대화문 등록 구글폼 페이지에서 이름, 연락처, 전자우편 주소, 성별, 연령, 거주지를 입력한 후 원시 데이터에 해당하는 대화 원문을 제출함
- 발화자의 카카오톡 대화명과 연락처를 함께 제출하게 하여 어노테이션에 활용함
- 구글폼에 입력한 정보들은 실시간으로 구글 스프레드시트에 응답 결과로 저장되어 관리자가 확인 및 데이터 가공이 가능함

[그림 Ⅲ-57] 온라인 서명 플랫폼을 이용한 전자 계약 과정 예시

2) AI 알고리즘 편향 방지 및 다양성 확보를 위한 데이터 획득 방안

- 본 과제를 통해 구축하는 대화 데이터는 일상대화와 토론 대화로서 주제의 다양성이 보장되어야 AI 알고리즘의 편향 방지에 유리
- 학습 시 과적합을 방지하기 위하여 소수의 크라우드 워커가 구축한 데이터의 전체 데이터의 대량을 차지하는 것을 방지하기 위하여 크라우드 워커별 작업 할당량 상한선을 적절히 정하여 반영함
- 데이터 검수 과정에서 화자 정보에 대한 정량 지표(참여자 수, 참여자 성별 및 연령대)를 활용하여 화자 특성에 따라 다양한 영역에 분포가 이루어지도록 함

3) 구글 폼(Google Forms)과 구글 스프레드시트를 활용한 데이터 정제

- 크라우드 워커는 제출한 대화 원문의 개인정보 및 민감 정보를 비식별화하여 제출함
- 크라우드 워커에게 개인정보 비식별화 지침을 제공하여 레이블링할 수 있게 함
- 구글폼에 입력한 정보들은 실시간으로 구글 스프레드시트에 응답 결과로 저장되어 관리자가 확인 및 데이터 가공이 가능함

2.4 획득 및 정제 기준

- 원시 데이터의 양품/불량
  - 모든 대화 참여자의 개인정보 수집 및 이용 동의와 저작권 활용 계약 체결이 이루어지지 않은 대화 데이터는 데이터셋에 포함될 수 없음
  - 모든 대화 참여자의 화자정보 제공이 이루어지지 않은 대화 데이터는 데이터셋에 포함하지 않음
  - 대화의 내용이 반사회적, 혐오적, 차별적, 선정적인 경우 데이터셋에 포함하지 않음
  - 대화의 내용이 텍스트에 비해 이모티콘, 사진, 동영상, URL 등 비텍스트적 요소가 지나치게 많은 비율을 차지하는 대화 데이터는 데이터셋에 포함하지 않음
- 개인정보 비식별화
  - 비식별화 지침을 참고하여 마스킹 비식별화를 수행하되 해당 위치에 어떤 내용이 있었는지를 라벨링함

[그림 Ⅲ-58] 구글 폼(Google Forms)을 이용한 원시 데이터 획득 예시

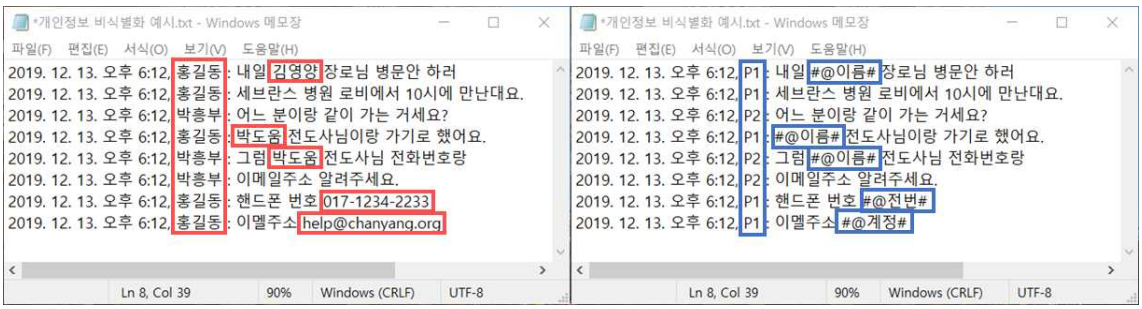
〈표 Ⅲ-52〉 개인정보 비식별화 유형 및 예시

언급내용		비식별화 (레이블링) 문구	예시	
유형	**항목**		**원문**	**비식별화(레이블링) 결과**
이름	실명	#@이름#	나는 홍길동이야	나는 #@이름#이야
	실명(변형)		길동구방구야 오늘 뭐해?	#@이름#야 오늘 뭐해?
	특수 애칭, 별명, 필명, 대화명			
	일반 애칭, 별명	대상 아님	자기야, 여보 등	변경 없음
	공인 실명		김연아, 빌게이츠 등	변경 없음
온라인	아이디	#@계정#	내 아이디는 sample로 찾으면 돼	내 아이디는 #@계정#로 찾으면 돼
	전자우편 주소(이메일)		sample@sample.com으로 보내	#@계정#으로 보내
	개인을 식별할 수 있는 URL	#@URL#	instagram.com/sample 내 인스타 주소야	#@URL# 내 인스타 주소야

언급내용		비식별화 (레이블링) 문구	예시	
유형	항목		원문	비식별화(레이블링) 결과
장소	상세 주소 (동 이하 주소만)	#@주소#	배송지는 서구 연희동 123로요	배송지는 서구 #@주소#요
	거주 아파트 및 건물명			
	거주지 역명 (지하철역, 기차역 등)	대상 아님	도곡역 3번출구로 오세요	변경 없음
	방문 장소 (비 정기적)		연세대 앞 정류장이야	변경 없음
	상호명		롯데리아에서 만날래?	변경 없음
각종 번호	고유식별 번호 (주민번호, 학번, 사번)	#@신원#	응 학번은 200101-1234567	응 학번은 #@신원#
	전화번호	#@전번#	언니 번호 010-1234-5678이야	언니 번호 #@전번#이야
	금융 번호 (계좌, 카드번호 등)	#@금융#	입금은 신한 110-234-456-789 홍길동	입금은 #@금융# #@이름#
	일련번호	#@번호#	사업자등록번호는 123-45-67890 입니다	사업자등록번호는 #@번호# 입니다
	(구매자)식별 번호			
	사업자 등록 번호			
	비밀번호			
출신 및 소속	출신 및 소속 학교	#@소속#	한국대학교에 재학중 입니다	#@소속#에 재학중 입니다
	출신 및 소속 직장			
	출신 및 소속 부대			
기타	비속어 등 위 보기에 없으나 비식별화, 또는 마스킹이 필요한 경우	#@기타#	-	-

- 이름 범주에서 실명을 비롯하여 화자의 특정이 가능한 실명의 변형, 대화명, 필명 등은 모두 비식별화 대상이며 일반적인 애칭이나 연예인, 공인 등의 이름은 비식별화 대상이 아님
- 온라인 범주에 속한 아이디, 전자우편 주소와 각종 번호 및 비밀번호 범주에 속한 주민등록번호, 운전면허번호, 전화번호, 통장계좌번호 등의 모든 번호는 비식별화 대상임
- 장소 범주의 정보들은 화자의 특정이 가능한 상세 주소와 아파트명 등은 비식별화 대상이며, 지하철역명, 상호 등은 비식별화 대상이 아님

- 출신 및 소속 범주에 속하는 학교, 직장 등의 정보는 모두 비식별화 대상임
- 제시된 기준에 포함되지 않더라도 다른 정보와 결합하여 화자의 특정이 가능하다고 판단되는 정보들은 모두 비식별화함
- '씨*', '존*'와 같이 사용 빈도가 높고 대표적인 유형의 욕설이나 비속어에 대해 우선적으로 마스킹 처리를 수행하고, 그 이외에 욕설이나 비속어의 마스킹도 순차적으로 진행함



[그림 Ⅲ-59] 개인정보 비식별화 예시

3 어노테이션/라벨링

3.1 어노테이션 / 라벨링 절차

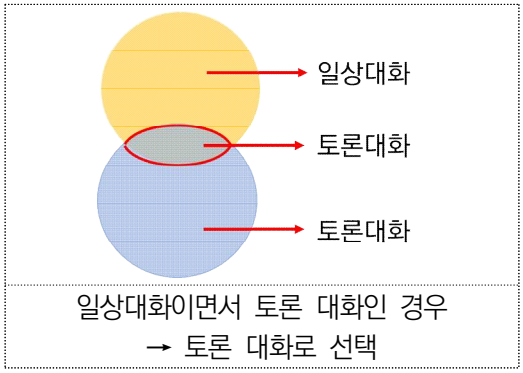
1) 데이터 형식 변환 및 화자 정보 결합

- 원시 데이터의 형식을 JSON 형식으로 변환함
- 카카오톡에서 추출한 대화 데이터는 카카오톡앱이 실행된 운영체제에 따라 일부 차이가 있으나 대체로 일정한 형식을 지니고 있으므로 자동화된 방법으로 형식의 변환이 가능함
- 형식 변환 시 대화 건별로 부가되어 제공되는 화자 정보와의 결합을 수행함

2) 대화 유형 분류

- 대화 유형은 다음의 두 가지 중 하나를 선택함

※ 일상대화: 2인 이상이 '일상'을 주제로 하는 대화
※ 도론대화: 하나의 토론 주제에 대해 대립되는 의견이 있는 대화



[그림 Ⅲ-60] 일상대화 토론대화 중복 시 토론대화 선택

〈표 Ⅲ-53〉 토론 대화문 예시

토론 대화문 예시	
2020년 9월 17일 오후 5:35, 김찍어 : 아 배고프다 오늘 중식이 땡기네 2020년 9월 17일 오후 5:35, 최부어 : 란쥬탕슉어때?? 2020년 9월 17일 오후 5:35, 김찍어 : 오 좋지좋지 거기 맛집이자나 안가본지도 오래당 2020년 9월 17일 오후 5:36, 최부어 : 응응 란쥬탕슉은 부먹기본이라 더 맘에 듬 ㅋㅋ 2020년 9월 17일 오후 5:36, 김찍어 : 엥 부먹기본이엇어? 난 찍먹이 좋아 부먹노노해 2020년 9월 17일 오후 5:37, 최부어 : 왜지?? 탕수육은 부먹이지! 튀김옷에 소스가 적절하게 스며들어야한다고! 2020년 9월 17일 오후 5:38, 김찍어 : 허참나~탕슉튀김의 바삭한 식감이 살려면 무조건 찍먹이지 소스가 너무 달면 조절할 수도 있구 2020년 9월 17일 오후 5:40, 최부어 : 아니지아니지, 소스가 폭신하게 스며든 튀김옷이랑 안에 고기랑 한번에 씹는게 가장 맛있는 조합이지 2020년 9월 17일 오후 5:42, 김찍어 : 하 이사람 정말 확고하네.. 그래 오늘은 설득당햇다치고 란쥬 가쟈 그치만 담엔 부먹이여 ㅋㅋㅋ	
설명	
대립 의견	탕수육 찍먹 vs 부먹

3) 대화 주제는 다음의 표를 참고하여 9개 항목 중 하나를 선택

- 주제 항목: 개인 및 관계, 주거와 생활, 상거래(쇼핑), 식음료, 여가 생활, 일과 직업, 행사, 미용과 건강, 시사/교육
- 주제 선택 요령: 육하원칙 중 '무엇을'에 해당하는 내용을 고려하여 주제를 선택

〈표 Ⅲ-53〉 대화의 주제 분류

항목	항목 예시(키워드)
개인 및 관계	이름, 전화번호, 가족, 국적, 고향, 성격, 외모, 개인의 기호(선호), 직업, 종교, 반려동물, 연애(관), 결혼(관), 이상형, 인간 관계, SNS
주거와 생활	숙소, 방, 가구, 침구, 주거비, 생활 편의 시설, 지역, 지리, 가전 제품, 자취, 잡안 일, 육아, 부동산, 주거시설, 이사, 생활비, 자동차, 날씨, 계절, 위치, 거리, 길, 이동 수단,

항목	항목 예시(키워드)
	이동 경로, 대중교통(지하철, 버스, 택시), 우편, 전화, 통신, 휴대전화, 인터넷 서비스, 은행, 관공서
상거래(쇼핑)	쇼핑 시설 및 장소, 식품, 의복, 가정용품, 물건 및 가격, 택배, 중고거래, 서비스, 교환 및 환불, 구매 후기
식음료	식사, 음식, 음료, 배달, 외식, 맛집, 식사 메뉴, 야식, 디저트, 요리
여가 생활	휴일, 취미, 동아리 및 동호회 활동, 관심사, 방학, 휴가, 행사, 술, 웹서핑, 문학, 음악, 미술, 공연, 전시, 스포츠 관람, 엔터테인먼트
일과 직업	취업, 스펙, 직장 생활, 업무, 회식, 급여, 계약, 협상, 회의
행사	초대, 방문, 소개팅, 약속, 가족 및 친척 행사, 공적 행사, 사적 모임(친목 모임), 여행 장소 및 경로, 여행 계획(일정, 숙소, 교통편, 여행 경비), 여행팁, 기념품, 여행사 및 여행 상품
미용과 건강	신체, 위생, 부상 및 질병, 치료 및 수술, 보험, 병원, 운동, 미용, 다이어트, 건강 검진, 약품 및 건강 보조 식품(용품)
시사/교육	학교 교육, 교과목, 진로, 학원, 진학, 입시, 시험, 자격증, 성적, 자기 계발, 외국어 학습, 스터디, 학문 및 학술 분야, 학회 및 세미나, 정체, 경제, 사회, 사건 및 사고, 법과 제도, 여론, 국제 관계, 재해 및 재난

3.2 어노테이션 / 라벨링 기준

- 어노테이션 종류
  - 단일 대화는 대화 메타 데이터를 포함하고 있는 헤더와 발화 텍스트를 포함하고 있는 본문으로 구성함
  - 헤더에 포함된 대화 메타 데이터는 대화 정보와 화자 정보로 구성함
  - 단일 대화는 대화 메타 데이터를 포함하고 있는 헤더와 대화 텍스트를 포함하고 있는 본문으로 구성함
  - 헤더에 포함된 메타 데이터에는 대화ID, 대화의 유형, 대화의 주제, 대화 참여자 수, 말차례 수, 발화 수, 발화자 정보가 포함됨
  - 본문은 개별 발화로 구성되며 각각의 발화는 발생 날짜와 시각, 발화ID, 발화 텍스트로 구성됨
  - 발화 텍스트가 매스킹된 정보를 포함하고 있을 때에는 마스킹을 플레이스홀더(place holder) 문자열로 대치함

3.3 어노테이션 / 라벨링 교육

- 관련 내용 없음

3.4 어노테이션 / 라벨링 도구 및 사용법

- 관련 내용 없음

4 데이터 검수

4.1 검수 절차

1) 품질 보증 활동

- 짧은 사업 기간 동안의 효율적인 학습 데이터셋 구축을 위하여 사업계획 단계부터 완료까지 사업 전반에 걸쳐 계획된 체계적인 프로세스에 의한 지속적 품질보증 활동을 통해 데이터의 품질이 계속 유지될 수 있도록 지원

2) 품질 관리 전략: 다단계 품질 검수를 수행

- 1단계: 기계적 검수
- 2단계: 중복 검수, 내용조건검수 등
- 3단계: 수작업 검수
- 4단계: 수작업 교차 검수
- 5단계: 전문가 검수
- 6단계: 외부 기관 품질 인증(TTA)

3) 품질 단계별 리스크 집중 관리

- 단계별로 영향을 미치지 않는 단계 내 리스크와 단계 간 리스크 구분 관리
- 샘플 데이터 생성을 통해 전 과정 작업 지침서 작성
- 검수와 피드백의 조기 반영과 재교육

4) 교차 검수에 의한 품질 유지

- 검수 단계에서 동일 데이터 항목에 대한 검수를 2인의 검수자가 독립적으로 수행하여 검수의 일관성과 품질을 유지함
- 2인의 검수자의 의견이 일치하지 않을 경우에는 검수책임자의 의견을 반영함

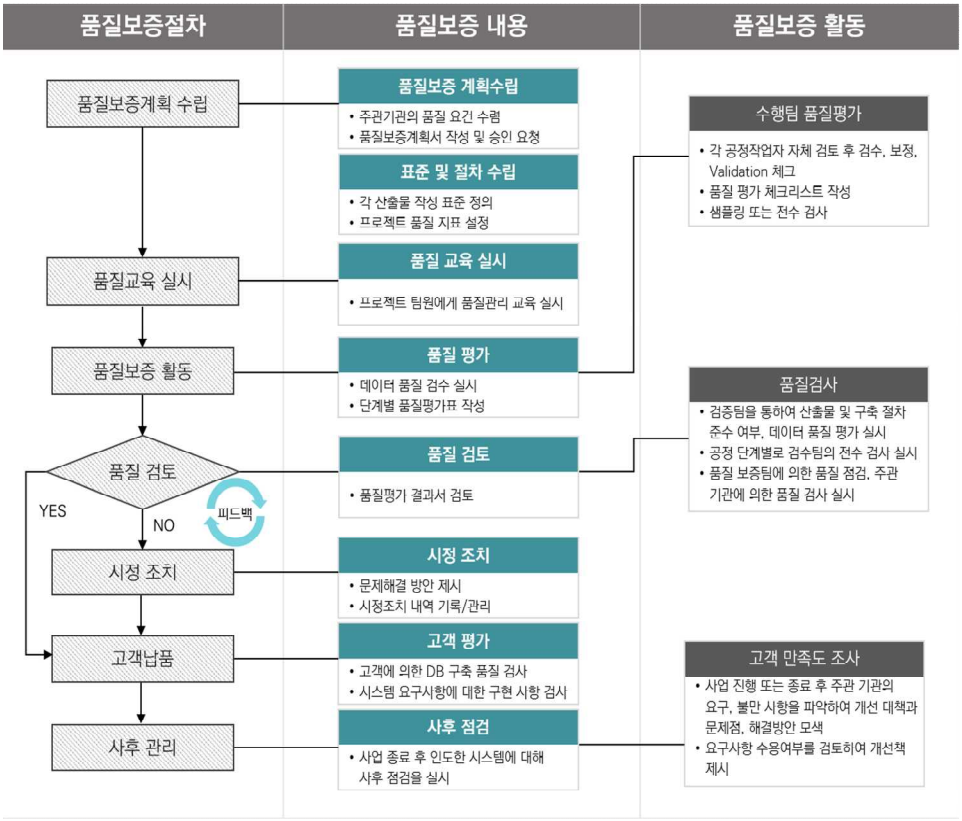
5) 대규모 언어 데이터 구축 경험 적용

- 수행기관 내 대규모 언어 데이터 구축 경험자 집중 투입
- 내부 검수 조직 구성 시 대학 및 관련 연구기관 언어 데이터 구축 경험자 채용

[그림 Ⅲ-61] 다단계 검수 절차

6) 데이터셋 검수

- 작성된 한국어 SNS 데이터셋에 대해 형식 및 내용적인 측면 검수 진행
- 검수 가이드를 기반으로 하여, 세부적인 검수 및 수정 사항의 경우 오타 확인, 질문 오류 확인, 답변 위치 확인, 질문에 특수문자 포함 확인 등을 통해 검수 진행



[그림 Ⅲ-62] 품질 보증 절차

4.2 검수 기준

1) 대화유형 검수

- 대화의 유형은 일상대화, 토론 대화로 분류
- 토론 대화는 하나의 토론 주제에 대해 두 개 이상의 대립되는 의견이 있는 대화를 적합한 대화로 정함

2) 대화주제 검수

- 대화의 주제 분류에는 개인 및 관계, 주거와 생활, 상거래(쇼핑), 식음료, 여가 생활, 일과 직업, 행사, 미용과 건강, 시사/교육의 9개 분야가 포함
- 대화의 주제는 육하원칙을 바탕으로 '무엇을'에 해당하는 분류가 선택되었는지 확인하여 판정
- 두 가지 이상의 주제에 포함될 수 있는 경우 더 주가 되는 주제 선택

- 주된 주제를 선택할 수 없을 만큼 뒤죽박죽인 대화의 경우에는 첫 번째로 나오는 주제 선택

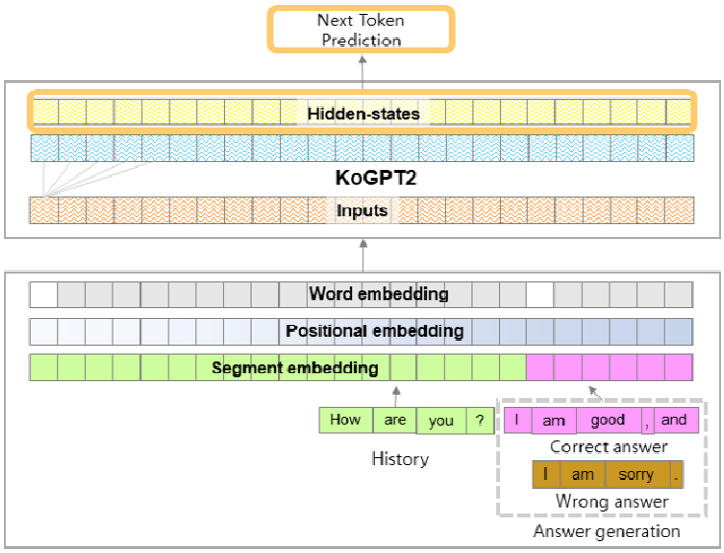
3) 검수 조직

- 품질 지원 조직
  - 효과적인 품질 관리를 위해 데이터셋 구축팀과 별개로 품질 관리 및 검수 전담 조직을 운영하며 전문 기술을 보유한 전담 인력을 투입하여 품질 관리 활동의 객관성 확보
- 품질 관리 조직 운영 원칙
  - 수행 조직 중 검수팀 및 품질 보증, 기술 지원 인력을 전문 인력으로 운영함
  - 검수팀 전담 인력은 데이터 품질 현황을 주기적으로 실무책임자에게 보고
  - 주기 또는 수시 품질 보증 활동을 실시하여 전체적인 품질 유지 및 관리

5 데이터 활용 방안

5.1 학습 모델

- 알고리즘 학습 방법
  - SK에서 한국어 위키 5,000,000 문장, 한국어 뉴스 120,000,000 문장, Other Corpus 27,400,000 문장을 학습하여 공개한 Ko-GPT2를 이용한 미세 조정을 수행함
  - KoGPT-2를 이용한 일생 대화 생성 모델의 개요를 그림으로 보이면 다음과 같음



[그림 Ⅲ-63] KoGPT-2를 이용한 일생 대화 생성 모델 개요

- 전체 구축 데이터의 50%인 100만건의 한국어 SNS 데이터를 8:1:1의 비율로 training/ validation/test 셋으로 분할하여 사용하였음
- 데이터 분할은 전제 학습 데이터 100만건의 주제 분포 비율을 반영하여 이루어졌음

● 신경망 모델 개발

- 대화의 맥락을 파악하기 위해 KoGPT-2 미세 조정 방법으로 개발 진행
- 화자2의 질문에 따른 화자1의 발화를 결정하는 것으로 개발 태스크 정의
- KoGPT-2 모델은 Transformer의 Decoder Layer로 구성된 단방향 모델로 대표적인 Auto Regressive 모델로서 주어진 맥락에 알맞은 다음 단어를 예측하는 형식으로 학습을 진행함
- 단방향 모델로 Next Token Prediction으로 학습된 것을 고려하여 미세 조정 진행
- 대화의 맥락을 파악할 수 있도록 대화 히스토리를 나열하고 마지막으로 화자2의 대한 화자1의 발화를 검색하는 Next Token Prediction으로 응답 생성을 진행하여 KoGPT-2를 미세 조정함
- 발화가 순차적으로 나열되기 때문에 대화 구간, 다음 발화 예측 구간을 Segment Embedding으로 구분하여 개발함

- 시범 모델의 학습 결과
  - 일상대화 생성 모델의 학습 성능과 데이터 유효성을 퍼플렉서티(perplexity)로 측정함
  - 퍼플렉서티는 주어진 확률 모델이 샘플을 얼마나 잘 예측하는가에 대한 측정 지표로 언어 모델을 평가하기 위한 평가 지표로 사용됨
  - 퍼플렉서티는 '헷갈리는 정도'로 의역할 수 있는데, 이는 모델이 테스트 데이터셋에 대하여 확률 분포를 얼마나 확실하게 예측할 수 있는지를 나타낸다고 할 수 있음. 그러므로 퍼플렉서티 점수가 높을수록 좋은 것이 아니라 낮을수록 (헷갈리는 경우가 적을수록) 좋다고 할 수 있음
  - 테스트 셋에 대한 평가는 Average_PPL 점수 값을 기준으로 판단함

5.2 서비스 활용 시나리오

- 배경 및 필요성
  - 코로나바이러스감염증-19 사태가 장기화되면서 1990년대 후반 이후 미국을 중심으로 시작된 '디지털 경제'로의 전환이 가속화됨
  - 디지털 경제의 핵심 요소 가운데 하나는 '비대면 경제'인바, 비대면 의사소통의 역할이 폭발적으로 증대하고 있음
  - 비대면 의사소통이 활성화됨에 따라 자연어 이해 및 생성 기술을 이용한 챗봇 서비스의 개발과 도입이 가속화되고 있음
  - 챗봇 서비스는 온라인 심리 구매 상담, 심리 상담 등 다양한 비대면 의사소통 환경에서 이용될 수 있으며, 특히 정해진 틀에 따라 진행되지 않는 비정형적 일상대화는 고품질 서비스의 구현이 어려우나 그 유용성이 매우 큰 것으로 사료됨
  - 대규모로 구축된 한국어 SNS 데이터는 일상대화 생성 기능을 갖춘 챗봇에 연구, 개발에 큰 역할을 할 것으로 기대됨
- 활용 사례
  - 일상대화 생성 서비스는 다양 형태의 응용 프로그램과 결합된 형태로 구현 가능함
  - 본 과제에서는 일상대화 생성 모델링 결과를 웹 API 형태로 이용할 수 있도록 제공하여 대화 요약 기능을 손쉽게 적용하고 시험할 수 있음
  - 일상대화 생성 API 규격은 다음과 같음

```
[9]: # 일상 대화 생성 API 사용

query = {
    'history': ['안녕'],
    'utterance': '만나서 반가워'
}
res = requests.post(url='http://nlplab.iptime.org:33244/dialog/kor',
json=query)
res.json()
```

```
[9]: {'result': ' 만나서 반가워. 너는 직업이 뭐니?'}
```

[그림 Ⅲ-64] 파이썬 언어의 requests 라이브러리를 이용한 일상대화 생성 API 활용 예

제6장 K-POP 안무영상 데이터

1 데이터 정보 요약

1.1 가이드 분류

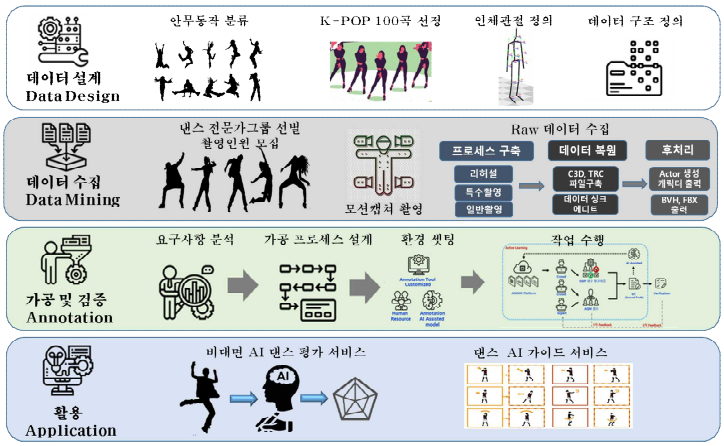
대분류	이미지	중분류	비전	소분류	JPG, PNG, MP4

1.2 데이터 정보

데이터 이름	K-POP 안무영상 데이터
데이터 요약	K-POP 커버댄스 모션캡쳐 촬영 대규모 안무영상 데이터 구축
데이터 출처	SMI 전문강사 및 댄서 프로파일 구성, K-POP 대표 100곡선정, 전수 모션캡쳐 스튜디오 촬영을 통한 데이터 구축

1.3 데이터 구축 개요

- POP 안무의 동작, 자세를 인식하는 인공지능 기술 및 응용서비스를 개발하기 위해 사람의 동작 별 자세와 더불어 안무 영상에 관한 데이터셋을 구축함



[그림 Ⅲ-65] POP 안무 영상 데이터 구축 개요

- 전체 개요
  - 본 과제는 데이터 설계, 수집, 가공, 검증, 활용의 단계로 이루어져 있음
- 데이터 설계
  - 기본 안무 동작 20개, K-POP 커버 댄스 100곡 안무 동작 등으로 동작 분류
  - 2D/3D 인체 데이터 및 자세 데이터 정의 및 포맷 설계
- 데이터 수집
  - 모션 캡처 카메라 멀티뷰 카메라 시스템으로 구성된 스튜디오 촬영
  - 200명 이상의 프로파일링 댄스 참여자에 대해 촬영 진행
  - 원천 데이터 획득 후, 데이터 수집 업체에서 촬영 데이터 품질검증 시행
- 데이터 가공
  - 인체 데이터 어노테이션 작업을 위한 저작 툴 제작
  - 크라우드 소싱 기반 데이터 가공
- 데이터 검증
  - 크라우드 소싱 기반 1차 데이터 검증
  - 내부 검수자 기반 2차 데이터 교차검증
- 데이터 활용
  - 비대면 AI 댄스 평가 서비스 개발
  - 댄스 AI 가이드 서비스 개발

1.4 구축 목적

- 전 세계적으로 유행하는 K-POP 댄스 동작 영상에 대한 데이터를 수집하여 인간의 동작, 자세, 관절의 움직임 정보를 정확하게 인식하고, 이와 연관된 산업 서비스를 개발하는데 필요한 인공지능 학습용 데이터셋을 구축

1.5 활용 분야

- 사람자세 분석, 홈피트니스, 교육, 실감미디어, 재활, 응용서비스 개발

1.6 유의 사항

- 전문 댄서와 일반인 댄서 (취미, 아마추어, 초보 등)를 대상으로 데이터를 수집
- 섭외된 전문 댄서들에게는 촬영된 본인의 영상 데이터의 노출범위와 사용범위에 대해 고지하고 초상권이용 동의서를 받음
- 본 컨소시엄의 구성 기관인 ㈜ SM Institute는 SM Entertainment 관계사로 서비스 활용목적에 대한 노래 및 댄스 인접저작권을 소유하고 있어 저작권 등의 문제를 해결함

2 데이터 획득 및 정제

2.1 원시 데이터 선정

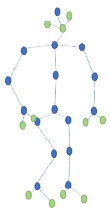
● 원시 데이터 선정

〈표 Ⅲ-54〉 K-POP 안무 영상 원시 데이터 선정

과제명	주요 내용	수집 방법	데이터 구축량	데이터 형식
K-POP 안무영상 데이터	K-POP 안무영상 데이터 구축	직접 촬영	40만 비디오 클립	- 2D 키포인트 - 3D 키포인트

● 원시 데이터 구성

〈표 Ⅲ-55〉 K-POP 안무 영상 원시 데이터 구성

	2D 영상	2D 인체 자세	3D 인체 자세
예시			
데이터 구성	댄서 촬영 이미지	29개 관절 위치	29개 관절 위치
포맷	png, jpg	json	json

- [2D 영상] 선정된 K-POP 100곡 안무 수행이 가능한 10대 20대 30대 200여명 댄서를 모집 전수 모션캡쳐 촬영을 수행함

- [2D 인체 자세] 모션캡쳐 센서 데이터와 카메라 파라미터를 이용한 2D 포즈데이터 추출
- [3D 인체 자세] 광학식 센서 기반의 3차원 인체 자세 수집을 통해 가공 과정의 휴먼 에러를 최소화 함

● 원시 데이터 수집 동작 분류

- Youtube 등 K-POP 댄스의 주요 접근 매체를 이용하는 대상들의 선호도 K-POP 100곡 선정
- 선정된 100곡 안무에 포함된 기본안무동작 20가지를 선정하여 동작 분류

〈표 Ⅲ-56〉 K-POP 100곡 안무분류

Index	대분류	소분류	곡 제목
1	K-POP 안무	K-POP 안무 001	Dancing King (엑소)
2		K-POP 안무 002	Love me Right (엑소)
3		K-POP 안무 003	Icecream cake (레드벨벳)
4		K-POP 안무 004	미인아 (슈퍼주니어)
5		K-POP 안무 005	Mr. Simple (슈퍼주니어)
6		K-POP 안무 006	Hurricane Venus (보아)
7		K-POP 안무 007	Only One (보아)
8		K-POP 안무 008	RingDingDong (샤이니)
9		K-POP 안무 009	Lucifer (샤이니)
10		K-POP 안무 010	Dream girl (샤이니)
11		K-POP 안무 011	Monster (엑소)
12		K-POP 안무 012	Call me baby (엑소)
13		K-POP 안무 013	Devil (슈퍼주니어)
14		K-POP 안무 014	I got a boy (소녀시대)
15		K-POP 안무 015	이것만은 알고가 (동방신기)
16		K-POP 안무 016	산소같은너 (샤이니)
17		K-POP 안무 017	View (샤이니)
18		K-POP 안무 018	상사병 (샤이니)
19		K-POP 안무 019	Cherry Bomb (NCT127)
20		K-POP 안무 020	Jopping (슈퍼엠)
21		K-POP 안무 021	Ko Ko Bob (엑소)

Index	대분류	소분류	곡 제목
22		K-POP 안무 022	Lotto (엑소)
23		K-POP 안무 023	Obsession (엑소)
24		K-POP 안무 024	Dumb Dumb (레드벨벳)
25		K-POP 안무 025	너같은 사람 또 없어 (슈퍼주니어)
26		K-POP 안무 026	누난너무예뻐 (샤이니)
27		K-POP 안무 027	영웅 (NCT127)
28		K-POP 안무 028	Super Human (NCT127)
29		K-POP 안무 029	소방차 (NCT127)
30		K-POP 안무 030	UN Village (백현)
31		K-POP 안무 031	4walls(fx)
32		K-POP 안무 032	Gee (소녀시대)
33		K-POP 안무 033	소원을 말해봐 (소녀시대)
34		K-POP 안무 034	Mr. Mr. (소녀시대)
35		K-POP 안무 035	Party (소녀시대)
36		K-POP 안무 036	Lion Heart (소녀시대)
37		K-POP 안무 037	My Name (보아)
38		K-POP 안무 038	Valenti (보아)
39		K-POP 안무 039	핫썸머(fx)
40		K-POP 안무 040	첫사랑니(fx)
41		K-POP 안무 041	빨간맛 (레드벨벳)
42		K-POP 안무 042	음파음파 (레드벨벳)
43		K-POP 안무 043	Run Devil Run (소녀시대)
44		K-POP 안무 044	Oh! (소녀시대)
45		K-POP 안무 045	다시만난세계 (소녀시대)
46		K-POP 안무 046	훗 (소녀시대)
47		K-POP 안무 047	Milky Way (보아)
48		K-POP 안무 048	운명 (동방신기)
49		K-POP 안무 049	Move (태민)
50		K-POP 안무 050	괴도 (태민)
51		K-POP 안무 051	Bad Boy (레드벨벳)
52		K-POP 안무 052	피카부 (레드벨벳)
53		K-POP 안무 053	몬스터 (레드벨벳아이린슬기

Index	대분류	소분류	곡 제목
54		K-POP 안무 054	Kissing You (소녀시대)
55		K-POP 안무 055	소녀시대 (소녀시대)
56		K-POP 안무 056	아틀란티스 소녀 (보아)
57		K-POP 안무 057	No. 1 (보아)
58		K-POP 안무 058	ID; Peace B (보아)
59		K-POP 안무 059	Listen to My Heart (보아)
60		K-POP 안무 060	Woman (보아)
61		K-POP 안무 061	평행선 (동방신기)
62		K-POP 안무 062	Something (동방신기)
63		K-POP 안무 063	Maximum (동방신기)
64		K-POP 안무 064	Rising Sun (동방신기)
65		K-POP 안무 065	왜 (동방신기)
66		K-POP 안무 066	쏘리쏘리 (슈퍼주니어)
67		K-POP 안무 067	Super Clap (슈퍼주니어)
68		K-POP 안무 068	이야이야오 (슈퍼주니어)
69		K-POP 안무 069	트윈스 (슈퍼주니어)
70		K-POP 안무 070	캔디 (백현)
71		K-POP 안무 071	으르렁 (엑소)
72		K-POP 안무 072	Love Shot (엑소)
73		K-POP 안무 073	늑대와 미녀 (엑소)
74		K-POP 안무 074	호랑이 (슈퍼엠)
75		K-POP 안무 075	100 (슈퍼엠)
76		K-POP 안무 076	Super Car (슈퍼엠)
77		K-POP 안무 077	Mama (엑소)
78		K-POP 안무 078	Halla(태티서)
79		K-POP 안무 079	아드레날린(태티서)
80		K-POP 안무 080	아름다워 (샤이니)
81		K-POP 안무 081	Why So Serious (샤이니)
82		K-POP 안무 082	펀치 (NCT127)
83		K-POP 안무 083	Regular (NCT127)
84		K-POP 안무 084	Touch (NCT127)
85		K-POP 안무 085	Once again (NCT127)

Index	대분류	소분류	곡 제목
86		K-POP 안무 086	Good thing (NCT127)
87		K-POP 안무 087	무한적아 (NCT127)
88		K-POP 안무 088	악몽 (NCT127)
89		K-POP 안무 089	Simon Says (NCT127)
90		K-POP 안무 090	Truth (동방신기)
91		K-POP 안무 091	Psycho (레드벨벳)
92		K-POP 안무 092	러시안룰렛 (레드벨벳)
93		K-POP 안무 093	Power Uo (레드벨벳)
94		K-POP 안무 094	짐살라빔 (레드벨벳)
95		K-POP 안무 095	RBB (레드벨벳)
96		K-POP 안무 096	Rookie (레드벨벳)
97		K-POP 안무 097	Miracle (슈퍼주니어)
98		K-POP 안무 098	The Boys (소녀시대)
99		K-POP 안무 099	Baby Baby (소녀시대)
100		K-POP 안무 100	Mirotic (동방신기)

〈표 Ⅲ-57〉 기본안무 20가지 분류

Index	동작 분류	동작	세부 동작	영문키워드
1	기본동작	아이솔레이션	아이솔레이션 머리	Isolation head
2			아이솔레이션 어깨	Isolation shoulder
3			아이솔레이션 가슴	Isolation chest
4			아이솔레이션 골반	Isolation hip
5		웨이브	웨이브 팔	Wave Arm
6			웨이브 몸통	Wave Body
7			웨이브 엉덩이	Wave hip
8		문워크	문워크	Moon Walk
9		바운스	바운스	Bounce
10		라이닝	라이닝 수평	Lining - Horizontal
11			라이닝 수직	Lining - Vertical
12			라이닝 대각	Lining - Diagonal
13		스텝	스텝, 앞과 뒤	Step-Front, Back
14			스텝 옆	Step-side
15		롤	롤	Roll
16		점프	점프	Jump
17		펀치	펀치	Punch
18		클랩	클랩	Clap
19		스톱앤고	스톱앤고(브레이크)	Stop and Go(Break)
20		팝	팝	Pop
21		턴	턴	Turn
22		킥	킥	Kick

2.2 규제관련 사항

- 원시 데이터 수집 고려사항
  - [개인정보 보호 준수] 법률적 확보방안 : 명확한 용도 제시와 사용 목적이 포함된 개인정보 및 초상권에 대한 사용 동의서를 작성하고, 데이터 제공자에게 충분히 알린 다음, 동의를 구하여 법률적인 문제를 해결할 예정임
  - [모든 데이터 직접 촬영] 웹 크롤링 같은 간접 데이터 수집이 아닌 컨소시엄 참여업체를 통한 직접 획득을 통해 원천 데이터 - 2D 영상 및 데이터를 확보할 예정임

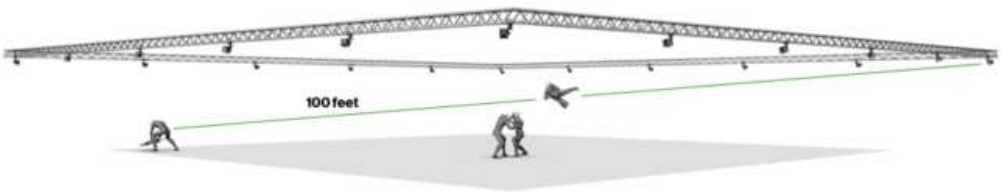
2.3 획득 및 정제 절차

1) K-POP 안무 동작 영상 데이터 수집주체 및 수집방안

- 단계별 댄스 동작 및 공통 시퀀스 동작 획득 등 원활한 인공지능 학습 데이터 구축을 목표로 함.
- 광학식 장비를 사용하는 촬영 스튜디오를 2개소 구성하여 병렬 촬영 진행

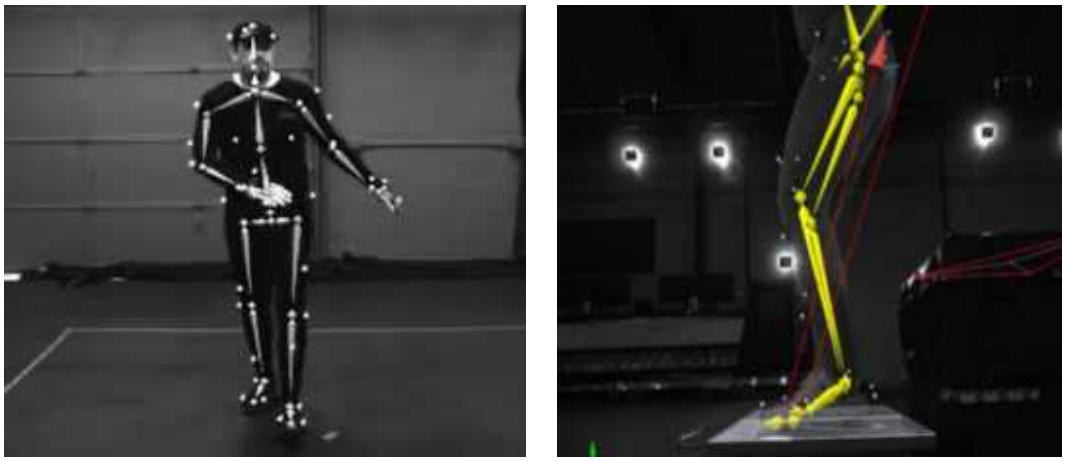
[그림 Ⅲ-66] 3D 모션 캡처 광학식 장비

- 고품질 모션 캡처 데이터 셋을 대량으로 확보하기 위해 광학식 카메라의 밀도를 늘려 마커 Swap이 나 Loss의 발생 빈도를 최소로 유지



[그림 Ⅲ-67] 3D 모션 캡처를 공간 환경 구성

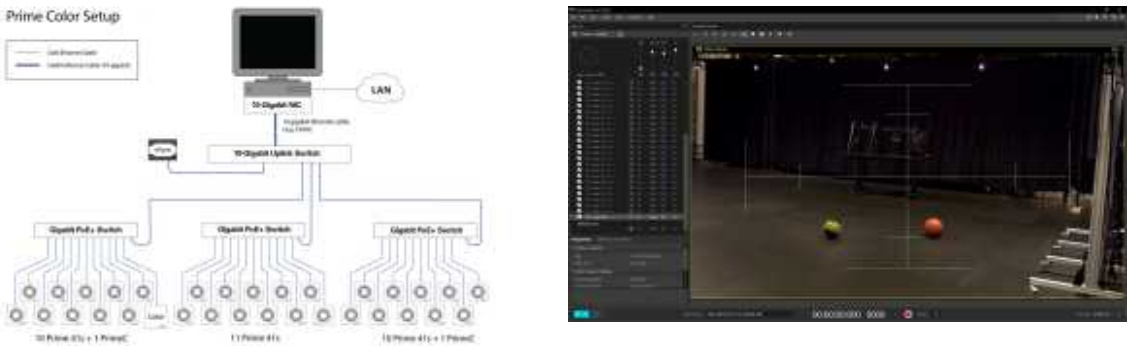
- 모션 캡처 데이터 셋의 품질을 높이기 위해 대상 액터의 신체 비율에 맞는 bvh 데이터를 통하여 3D 관절 데이터 수집



[그림 Ⅲ-68] 3D 관절 데이터

2) 데이터 수집 주체

- 고해상도 고속 다시점 카메라를 활용하여 (성별, 나이, 신체조건, 템포 등 유형 통제) 구성된 시나리오를 바탕으로 유니토 엔터테인먼트에서 촬영 및 가공, 학습용 안무 영상 10만 개 수집 (3D 관절 정보와 연동)



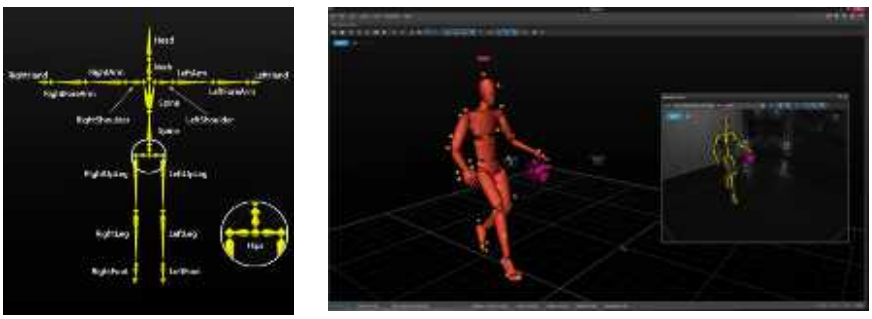
[그림 Ⅲ-69] 고해상도 고속 다시점 카메라 활용한 데이터 수집

- 모션 캡처 장비를 활용하여 (성별, 나이, 신체조건, 템포 등 유형 통제) 구성된 시나리오를 바탕으로 촬영 및 가공, 학습용 3D 데이터 및 3D 관절 정보 10만 개 구축 (고품질 안무 영상과 연동)

[그림 Ⅲ-70] 모션 캡처를 통한 3D 관절 위치 데이터 수집 과정

3) 데이터 수집 준비

- K-POP 안무가 가능한 연습생, K-POP 안무 수강생 또는 촬영 지원자를 대상으로 공연자를 섭외함
- 개인정보 수집 이용 동의한 학습용 데이터 구축을 위한 공연자 (SM Institute) 섭외 및 안무연습
- 최종 수집되는 모션 데이터의 정확도를 높이기 위하여 공연자 개개인별 BVH 포맷 생성 및 라벨링



[그림 Ⅲ-71] 공연자 개개인별 BVH 포맷 생성 및 라벨링

- 모션 캡처 시스템 운용, 고해상도 카메라 기반 촬영 진행

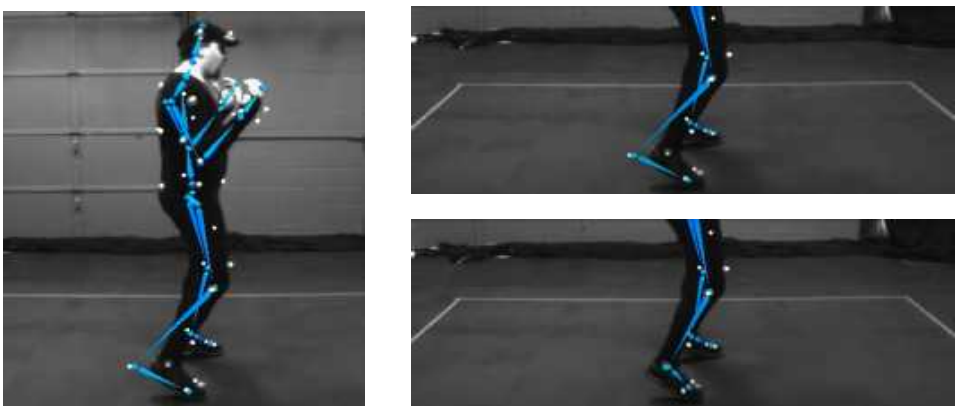
[그림 Ⅲ-72] 모션 캡처 시스템 및 고해상도 카메라 기반 촬영의 진행

4) 데이터 수집 방법

- 학습용 데이터 구축을 위한 공연자의 섭외, 개인정보수집이용 동의 절차 진행 및 배우 의상/모션 캡처 시스템 운용, 고해상도 다시점 카메라 (DSLR) 기반 촬영 진행
- 공연자의 머리와 의상 및 배경은 형태의 원활한 복원 품질을 위하여 통제가 필요하며, 단체인 경우 대형의 이동이나 공연자 간의 동작 가림이 최소화될 수 있는 촬영 구도 필요

5) 촬영시 데이터 정제 방안

- 데이터셋 촬영 시 발생하는 데이터 오류를 실시간 영상으로 식별하여 후처리 프로그램 등을 통하여 보정



[그림 Ⅲ-73] 데이터 오류의 후처리 프로그램 보정

2.4 획득 및 정제 기준

1) 정제기준

- c3d데이터는 매 프레임을 확인하여 occlusion이나 swap의 발생여부를 확인하여 보정한다
- 보정 완료된 c3d로 생성된 bvh는 별도의 담당자가 전체 프레임을 확인 후 공유한다
- 영상데이터는 모션캡쳐 데이터와의 동기화여부를 확인하여 후반작업자에게 공유한다

2) 중복성 방지

- 안무영상의 경우 전면과 측면등 최소한의 각도에서 촬영된 영상만 사용하여 중복을 방지한다

3) 비식별화

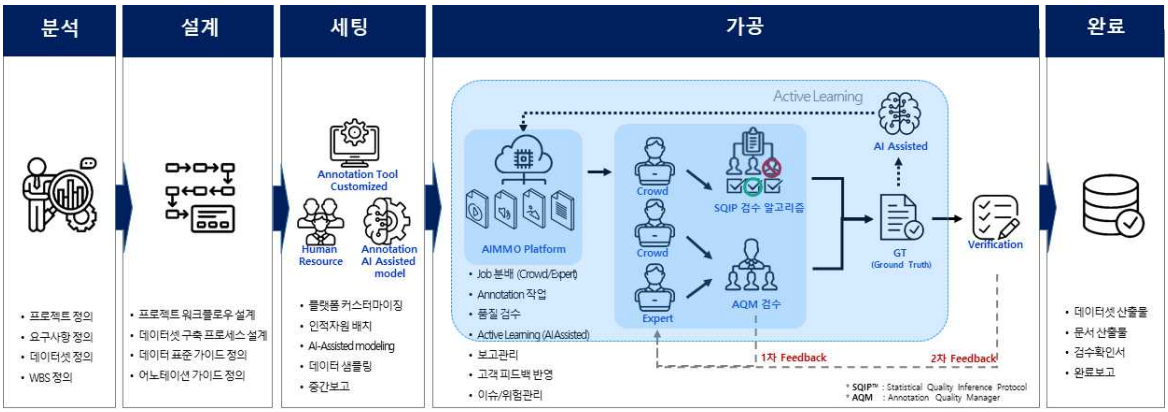
- 모션캡쳐 데이터의 경우는 비식별화 의제가 발생하지 않으며 안무영상의 경우는 초상권과 관련하여 개별 사용동의를 받아 수집한다

3 어노테이션/라벨링

3.1 어노테이션 / 라벨링 절차

● **(어노테이션/라벨링 절차)**

1) 라벨링 절차를 '분석', '설계', '세팅', '가공', '완료' 단계로 세분화 하여 추진

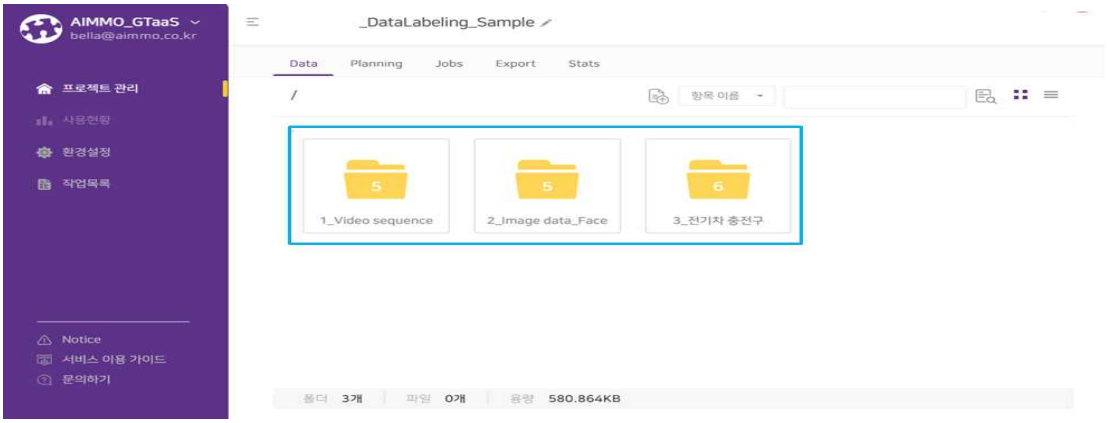


[그림 Ⅲ-74] 학습용 데이터 가공 프로세스

2) (분석) 프로젝트 정의, 요구사항, 정의, 데이터셋 정의

3) (설계/세팅) 프로젝트 Workflow를 하위 작업 단계(Stage)로 세분화하여 설정

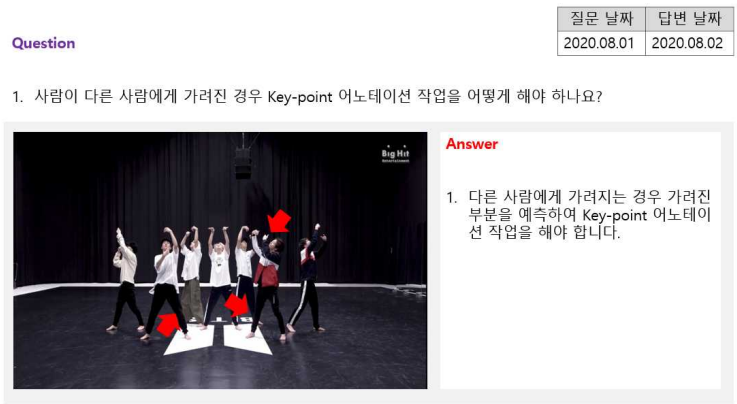
- 프로젝트 조직 설정 및 이메일을 통한 작업자 초대
- 프로젝트 어노테이션 유형, 분류 기준을 설정
- 프로젝트 작업 이미지 포팅



[그림 Ⅲ-75] 프로젝트 설계 화면

4) (가공/완료) 작업자는 온라인(오프라인 가능) 환경 어디서나 가공 작업이 가능하며, 작업이 완료되면 검수자에게 자동 전달

- 작업 결과를 실시간 확인하여 재작업 요청이 가능하며, 플랫폼내 검수자-작업자간 실시간 Q&A 소통, 가이드라인 기준 보완 등 진행
- 이해관계자(고객, 관리자 검수자, 작업자)는 가이드 기준을 상시 공유하여 일관성 유지



[그림 Ⅲ-76] 검수자-작업자간 Q&A 소통

3.2 어노테이션 / 라벨링 기준

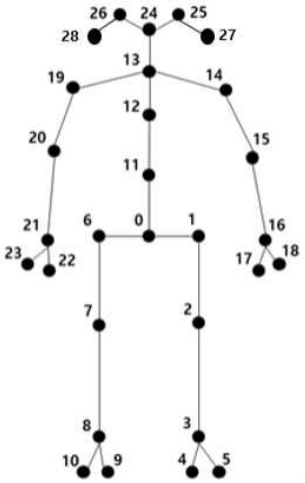
1) 바운딩 박스 데이터

- 화면 내 사람의 실루엣이 박스를 벗어나지 않는 형태로 표현

2) 사람 동작 관절 데이터

- 인체를 움직이는 관절을 기준으로 29개 주요 포인트를 선정하고 이들에 대한 키포인트(Key-point) 어노테이션 수행

Index	관절정보	Index	관절정보
0	가운데 엉덩이	15	왼쪽 팔꿈치
1	왼쪽 엉덩이	16	왼쪽 손목
2	왼쪽 무릎	17	왼쪽 손바닥 엄지
3	왼쪽 발목	18	왼쪽 손바닥 약지
4	왼쪽 엄지발가락	19	오른쪽 어깨
5	왼쪽 새끼발가락	20	오른쪽 발꿈치
6	오른쪽 엉덩이	21	오른쪽 손목
7	오른쪽 무릎	22	오른쪽 손바닥 엄지
8	오른쪽 발목	23	오른쪽 손바닥 약지
9	오른쪽 엄지발가락	24	코
10	오른쪽 새끼발가락	25	왼쪽 눈
11	허리	26	오른쪽 눈
12	가슴	27	왼쪽 귀
13	목	28	오른쪽 귀
14	왼쪽어깨		



[그림 Ⅲ-77] 사람 동작 관절 데이터 키포인트 정보

3.3 어노테이션 / 라벨링 교육

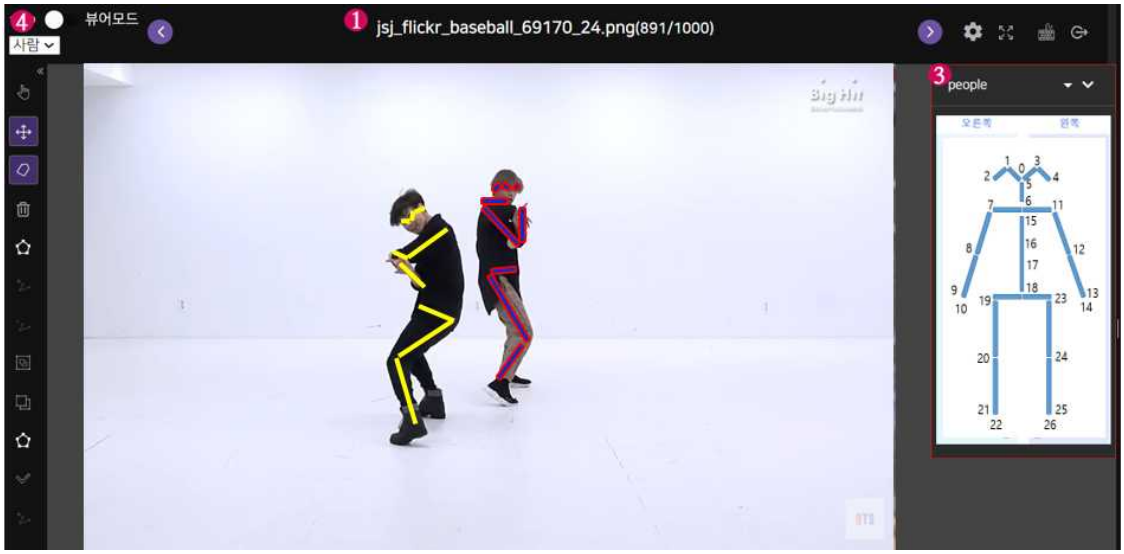
- 크라우드 작업자 난이도별(초급/중급/고급) 온/오프라인 교육 실시
- 교육 및 작업 결과에 따라 프로젝트 수행 후 검수자, 내부 관리자로 채용 기회 제공

3.4 어노테이션 / 라벨링 도구 및 사용법

1) 어노테이션 전문기업(에이모)이 개발한 저작도구 활용

2) (작업) 작업자 활용을 위한 편의 기능 탑재(작업이동, 화면 조절 등)

- (작업 이동) 작업後, 작업中, 작업前 파일을 간편을 간편하게 이동
- (화면 조절) 화면 상단에서 작업 화면 크기/밝기/색상, 실행 취소, 재실행 등 제공
- (가공 지원) 어노테이션 작업 편의를 위한 간편하게 클릭만으로 이미지 이동, 바운딩박스, 라벨, 선택 삭제, 초기화 기능 등을 제공



[그림 Ⅲ-78] 데이터 저작도구 활용 화면

3) (결과물) 작업 결과물을 다양한 형식 제공할 수 있도록 데이터 포맷 컨버터 기능 지원

4) (저작도구 공개) 과제를 통해 진행된 저작도구는 소스와 기술 매뉴얼(데이터셋 형태, 규모, 특성 등)을 공개하여 외부에서 활용이 가능

4 데이터 검수

4.1 검수 절차

- 검수를 위해 어노테이션 공정별 6단계 절차로 검수 진행(2인 이상 수작업 검증 실시)
  - 공정별 품질관리 목표를 설정하고 진단·개선을 통해 고품질 AI 학습용 데이터 제작

〈표 Ⅲ-58〉 데이터 품질관리 프로세스

구 분	프로세스	설명
데이터 분석	대상 식별	• 고객사의 품질관리 요구사항을 확인 • 품질관리를 수행할 대상을 구체화 및 문서화
▼		▼
데이터 설계	규칙 정의	• 품질관리 대상에 대한 프로파일링을 시행하고 • 품질 측정 및 통제를 위한 지표를 설정 • 설정된 품질규칙은 데이터 가공 업무규칙에 반영
▼		▼
데이터 가공	측정	• 데이터 가공 결과물 중 품질관리 대상에 대한 품질 측정
	분석	• 품질 측정 결과를 품질지표와 비교하여 시사점 도출 • 개선이 필요한 부분에 대한 원인 및 개선방법 분석
	개선	• 오류의 영향도 및 시급성을 고려하여 개선 시행
	통제	• 품질측정-분석-개선이 선순환 구조를 이룰 수 있도록 지속적인 모니터링 수행

4.2 검수 기준

〈표 Ⅲ-59〉 정확도 및 유효성 측정 지표/기준

구분		측정 지표	정량 목표
정확도	구조 및 형식	어노테이션 포맷 정확도	정합률 97% 이상
	참값(Ground Truth)	참값 정확도	오태깅율 3% 이하
유효성	학습 성능	관절 위치	AP 기준 0.65 이상

● TTA 품질 기준

- (TTA 품질검증) 학습데이터 품질검증에 필요한 자료 및 환경(도구) 제공 및 적극 지원

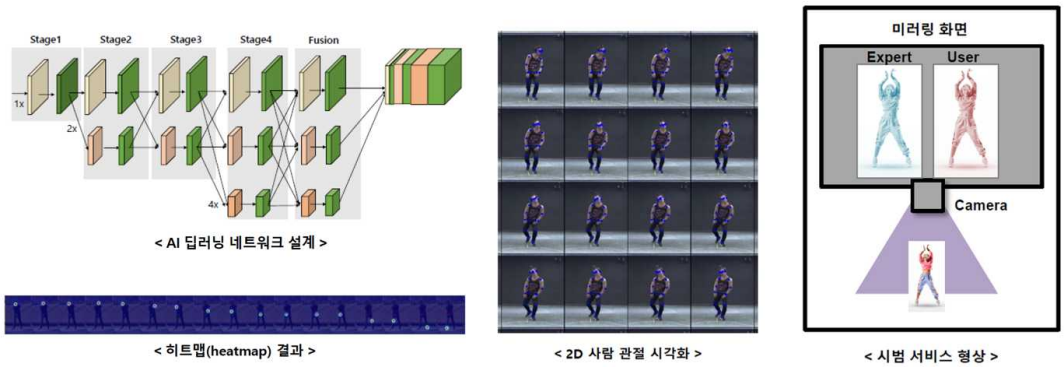
〈표 Ⅲ-60〉 TTA 품질검증 체계별 지원사항

구분	구축공정	정확도	유효성
검증대상	공정 전주기	데이터 및 저장소	학습모델
검증방법	• 문서 검토 • 수행기업 인터뷰 • 현장 점검, 자료 확인	• 전수 또는 샘플링 검사 • 자동화 검수 도구 • 검증 데이터 분석	• 학습조건 설정 및 수행 (데이터 구분, 반복 횟수 등)
지원사항	• 작업공정도 제공 • 인터뷰/현장점검 지원	• 품질검증 데이터 제공 • 저작도구 활용 지원	• AI 학습결과 검증 지원
확인	품질검증 결과서(구축공정, 정확도, 유효성)		

5 데이터 활용 방안

5.1 학습 모델

- 2D Human Pose Estimation AI 모델
  - 구축 데이터기반 2D 자세추론 AI 모델설계
  - 고해상도/저해상도 정보 포함 HR-Net 기반 네트워크 구성, 관절위치 고해상도 히트맵 결정
  - 서비스는 사용자의 관절좌표 확인 및 안무배우기를 체험하는 형태로 개발



[그림 Ⅲ-79] 2D Human Pose Estimation AI 모델

- AI 모델 학습 방법(안)

〈표 Ⅲ-61〉 AI 모델 학습 방법

AI 모델 학습 방법		
개발 언어	Python	
프레임워크	Pytorch	
학습 네트워크 개념	CNN 기반의 지도 학습	
학습 데이터	Training Data set	약 40 만장
	Vaildation Data set	약 4 만장
	Test Data set	약 4 만장
학습 조건	epoch = 200, batch = 64, iteration 등	
검증/평가 방법	k-fold cross validation (k=5)	

5.2 서비스 활용 시나리오

- K-POP 안무영상 데이터를 활용한 응용서비스는 댄스체험 및 AI평가 기능의 시범서비스 앱 다운로드 형태로 제공될 예정이다. 앱을 AI Hub 시범서비스 페이지를 통해 다운로드 받고 설치한후 아래와 같이 K-POP 안무체험 서비스가 실행되고, 각 기능을 체험한 후 평가결과를 알고리즘이 자동으로 알려준다. 아래 그림은 시범서비스 진입 예시화면을 보여준다

[그림 Ⅲ-80] 댄스체험 및 AI 평가 기능 시범 서비스 앱 화면

- 아래 그림은 댄스 체험화면 예시를 보여준다. 외쪽은 강사 영상이 표시되고, 우측은 실제 사용자의 카메라 입력 화면이 표시되어 댄스 체험을 할 수 있도록 설계되었다. 우측 하단바의 댄스 동작 결과에 대한 표시 시각화 기능을 통해 자동으로 댄스평가를 사용자가 인지할 수 있도록 구성될 예정이다

[그림 Ⅲ-81] 댄스체험 및 AI 평가 기능 시험 서비스 앱 실행 화면

제7장 고해상도 LF AI 학습용 데이터

1 데이터 정보 요약

1.1 가이드 분류

대분류	이미지	중분류	비전	소분류	PNG

1.2 데이터 정보

데이터 이름	고해상도 LF AI 학습용 데이터
활용 분야	• 디스플레이 제조사 • 다시점 기반 영상 해석 • VR/XR 방송 및 공연 • 3D, 홀로그램 프린팅
데이터 요약	• 실내 및 실외 25 시점 • 2K(1080p), 가변 fps(1 ~ 24fps) • 10초당 h264파일 7개 동시 생성. • 디코딩 후에는 PNG, JPG로 구성 • 어노테이션 정보 : Json 파일

1.3 데이터 구축 개요

〈표 Ⅲ-62〉 고해상도 LF AI 학습용 데이터 구축 개요

중항목	소항목	내용	결과물
데이터 획득	데이터 수집	• 다양한 조건으로 15개 이상의 위치에서 100시간으로 촬영된 고해상도 LF 원천데이터(50TB 이내) - 촬영 위치 및 조건: 실내·외, 다양한 시간대 - 촬영대상: 자연물 및 인공물 - 촬영 시간: 100시간	고해상도 LF 원천데이터
	데이터 정제	• 고해상도 LF AI 학습용 데이터 생성을 위한 데이터 정제작업	고해상도 LF 정제 데이터

중항목	소항목	내용	결과물
		- 왜곡, 노이즈 촬영 실수 등 품질 이상 제거 - 과도한 모션 블러 제거 - 중복데이터 제거 - 데이터 분류 - 개인정보 및 프라이버시 문제 제거	
데이터 가공	어노테이션	• 고해상도 LF AI 학습용 데이터 생성을 위한 어노테이션 - Object ID - Bounding-Box - Mask - 객체 깊이 정보 - 객체이동 정보	고해상도 LF 어노테이션
	데이터 생성 자동화	• 고해상도 LF AI 학습용 데이터 생성을 위한 어노테이션 - GPU 가속 어노테이션 정보 추출 - 1차 자동 다중객체 인지, 분류 및 레이블 생성 - Positive/Negative stroke 입력 자동 생성 - 상용 Cloud platform 활용	고해상도 LF 초기데이터 자동화 방법
	저작도구 및 데이터 저작	• 전처리 데이터를 데이터 가공자들이 직관적으로 짧은 교육과정 내에 편집 업무를 수행할 수 있는 저작도구를 개발함 - 바운딩 박스 편집 - 마스킹 편집 - 수동 레이블링	고해상도 LF 초기데이터 저작도구 및 데이터 저작
	품질관리	• 카메라 캘리브레이션 기반 품질관리 - 멀티 카메라 어레이 카메라 캘리브레이션 - 고해상도 LF AI 데이터 원천데이터 품질관리	고해상도 LF 데이터 품질관리
데이터 활용	AI 응용서비스	• 고해상도 LF 데이터를 이용한 동작 인식 • 고해상도 Light field 데이터를 활용한 효율적 객체 분리 서비스	• 고해상도 LF 데이터 동작 인식 서비스 • Web 기반 고해상도 Light field 데이터 객체 분리 서비스

1.4 구축 목적

- 고해상도 Lightfield 카메라에서 생성되는 다차원 이미지 집합에서 개별 이미지들의 정보를 활용하여 Lightfield 데이터를 구축하여 ① 자유 시점 이미지 시퀀스, ② 재초점, 관심 영역의 포커스 블러 및 디블러, ③ 깊이 정보 추출뿐 아니라 영상처리 및

이해 연구 주제인 ④ 객체 레이블링 및 마스킹, ⑤ 객체 인식, 6) 객체이동 경로 추적 및 추정 등 다양한 목적에 활용할 수 있음

- 본 사업으로 구축된 데이터를 활용하여 데이터 이용자는 ① 영상 속의 객체를 인식하고, ② 영상 속 객체의 이동 방향을 추정하고, ③ 영상 속 객체의 깊이 정보를 획득하고 ④ 다시점·다초점 영상 정보를 통한 재초점, 자유 시점 생성 등의 다양한 응용이 가능하여 엔터테인먼트 업계 뿐 아니라 다양한 공학적 용도로 활용할 수 있음
- 구축되는 LF AI 학습용 데이터를 활용하여 객체추적, 객체 인식과도 같은 전통적인 이미지 프로세싱 주제와 함께 영상기반 콘텐츠 제작에 필수적인 영상합성을 위한 깊이 정보 추출 및 4D 인터렉티브 콘텐츠 제작에 필수적인 다시점 객체 복원 및 생성, 딥러닝을 활용한 객체 복원 및 동작 추정 분야의 품질 향상을 위해 폭넓게 활용될 수 있음
- 다시점·다초점 영상데이터가 국가적으로 구축되지 않은 상황이기 때문에 다른 영상기반 구축사업의 AI 학습용 데이터와 혼용하여 사용되거나 다른 데이터셋을 이용한 AI 엔진의 성능 검증 혹은 성능 개선 등에 다양하게 활용 가능함
- 제안 컨소시엄은 국내의 학계/산업계의 연구/개발자들이 마음껏 사용할 수 있도록 지재권이 해결된 원천데이터를 확보하고, 정교하게 가공/ 공개하여 다양한 시점에서의 정교한 객체 인식, 객체추적, 깊이 정보 추출이 가능할 수 있는 인공지능 학습용 데이터를 개발할 수 있도록 하고자 함

1.5 활용 분야

- 다시점, 다초점 데이터를 통해서 동작 인식, 영상 세그맨테이션, 객체추적 기술에 다양하게 활용될 수 있음
- 또한, 다시점, 다초점 데이터를 활용하여 Volumetric 영상제작에서 Deep learning을 활용한 어플리케이션을 보다 고도화할 수 있음
- Volumetric 영상제작에서의 Deep Learning 활용은 기존의 컴퓨터 비전 기술에 국한하여 다시점 영상으로부터 3D를 복원하는 방식이 아닌 미리 저장해둔 3D 동작 및 객체 DB로부터 다시점 영상에서 취득된 자료를 학습하여 가장 높은 확률의 객체를 추정하여 이를 3차원 상에서 복원하는 방식임
- 이러한 연구 분야들 외에도 전통적인 이미지 기반의 객체 인식, 추적, 분리 등에 활용될 수 있는 데이터이므로 활용도가 높음

- 단순 흉부영상, 흉부 전산단층촬영 영상, PET CT 영상 판독 기술을 종합적으로 판단하여 최종적으로 악성의 예측도를 제고하는 프로그램 개발 및 실증 데이터로 활용
- 폐암의 진단 및 예후 예측을 위한 인공지공 기반 예측 요인(feature) 개발을 위한 실증사업 솔루션 개발 및 사업화 확장

1.6 유의 사항

- 영상 데이터에 포함되는 인물에 대하여 계약을 통하여 저작권을 획득하는 것이 가장 확실한 방법. 하지만 촬영 현장에서의 제약사항을 고려하면 저작권을 획득한 인물외의 다른 인물(개인) 또는 다른 인물(개인)과 연관되어 개인의 식별을 가능하게 하는 객체(자동차, 소지품 등)가 포함될 가능성을 배제할 수 없음. 이에 저작권의 획득은 계약을 통한 저작권획득, 촬영현장(로케이션)에서의 통제를 포함해야 함
- 촬영현장을 통제하고 고지한다 하더라고 포함되는 인물이나 연관된 객체가 있다면 촬영 직후의 영상 데이터는 개인정보의 침해가능성을 포함할 수 있어 이러한 영상정보의 안전 운반에 대하여 고려해야 함
- 촬영된 영상 데이터를 기준으로 기술적으로 비식별 처리를 해야 될 대상을 정의하고, 대별로 비식별 처리방안을 고려해야 함. 기술적 처리를 위한 비식별의 대상은 아래 항목을 포함하여 고려해야 함
- 결합을 통하여 식별 가능한 정보 : 여러 조합으로 개인의 식별이 가능하게 하는 객체의 영상정보로 특히 아래 객체들은 주의해서 기술처리가 필요함

2 데이터 획득 및 정제

2.1 원시 데이터 선정

〈표 Ⅲ-63〉 원시데이터 선정 - 사무실

Scene	S#1		
장소	사무실	**특정 예시**	위지윅 스튜디오 별관 사무실
선정 이유	  〈드라마 '미생', '변혁의 사랑'의 사무실 장면〉 • 사무실 내부에는 분류하기 좋은 다양한 객체(모니터, 컵, 의자, 사람, 책상, 사무용품 등)가 배치되어있음. • 영화나 드라마 중 사무실 장면의 등장이 많기에 활용도가 높을 것으로 예상됨 • 특히 하나의 사무실 안에 극중에 등장하는 배우들이 자리에 맞게 곳곳에 위치되어 각각의 연기를 진행하는 구조이므로, 각 부분에 초점과 시점을 맞춰 여러 번 촬영이 진행되는 것보다, 촬영이 된 결과에서 다각도에서 초점을 바꿔가면서 postprocessing을 할 수 있다면 촬영의 효율성이 높아질 것으로 예상됨		
촬영 방법	• 내부의 사람들에게 촬영 동의를 얻은 후에, 최소 3명 이상이 근무 중인 실제 사무실 내부를 촬영한다. • 움직임이 많을 시간대에 한 자리에 카메라를 두고 촬영 • 움직임이 적을 시간대에 카메라를 움직이며 촬영		
촬영 분량	16시간		

〈표 Ⅲ-64〉 원시데이터 선정 – 모델하우스

Scene	S#3		
장소	모델하우스	특정 예시	근처 모델하우스 또는 이케아
선정 이유	  〈모델하우스와 드라마 '부부의 세계' 예시〉 • 모델하우스 내에 실내 객체가 매우 많아 다중객체 분석에 유용한 데이터를 획득할 수 있음 • 주거공간의 경우 소규모의 실내에 많은 객체가 들어와 있는 구조이기 때문에, 촬영의 각도에 따라 보이는 부분의 차이가 큼 • 원경과 근경에 대한 관심사가 다를 수 있으므로 초점의 변경 또한 필요한 장면을 취득할 수 있음 • 영상 사업에서도 주거환경을 배경으로 한 작업들이 많으므로, 주거 환경 내의 배우의 모습을 다각도로 촬영할 수 있다면 많은 도움이 될 것이라 생각됨.		
촬영 방법	• 모델하우스 촬영 허가를 받은 후 촬영 진행 • 내부의 개체만 촬영하거나, 사람과(배우나 촬영 스태프) 함께 촬영 • 카메라를 조금씩 움직이며 촬영 • 카메라를 한곳에 두고 사람의 움직임을 촬영 • 사람의 움직임은 데이터를 활용하기 좋도록 미리 동선을 정한 후 촬영		
촬영 분량	4시간		

〈표 Ⅲ-65〉 원시데이터 선정 – 카페

<table>
<tr><td>Scene</td><td colspan="3">S#6</td></tr>
<tr><td>장소</td><td>카페</td><td>특정 예시</td><td>성수동 카페</td></tr>
<tr><td>선정 이유</td><td colspan="3"> 

〈카페와 드라마 ‘도깨비’의 카페촬영 예시〉

• 카페의 경우 창이 크게 나 있는 경우가 많아 조명 조건 및 원경, 근경 촬영이 활용도가 높음
• 또한 카페는 국내 드라마와 영화, 예능 등의 장면에서도 자주 등장함
• 이 부분을 촬영 데이터 학습을 통해 카페와 같은 원경/근경이 동시 존재하는 장소에서의 다시점 다초점 촬영 데이터가 유용할 것으로 예상됨</td></tr>
<tr><td>촬영 방법</td><td colspan="3">• 창이 크게 나있는 카페를 선정하여 섭외.
• 배우 2명 이상을 섭외하여 창 앞의 테이블에 앉아 음료와 음식을 마시며 대화 연기를 하는 것을 요청.
• 카메라를 한곳에 두고 배우의 대화를 촬영.
• 드라마의 장면처럼 한 사람의 모습을 zoom 하는 등의 카메라를 움직이는 촬영을 포함.</td></tr>
<tr><td>촬영 분량</td><td colspan="3">2시간</td></tr>
</table>

〈표 Ⅲ-66〉 원시데이터 선정 – 공원

Scene	S#13		
장소	공원	특정 예시	반포 한강 잔디공원
선정 이유	  〈한강 잔디공원과 예능 '하트시그널' 예시〉 • 서울에서 자연 배경을 촬영할 수 있는 공간임 • 그러한 장점으로 예능에서 자연 배경이 필요한 외부 촬영 시 자주 활용됨 • 다시점 다초점의 촬영을 통해 자연 요소 안의 물체의 시점 변화에 대한 데이터를 획득할 수 있음. • 실외 객체에 대한 자연스러운 데이터를 획득할 수 있으리가 기대됨		
촬영 방법	• 배우를 섭외하여 다양한 장면 연출 • 텐트, 캠핑 의자 등의 객체를 구비하여 다양성을 확보. • 운동이나 스케이트보드, 자전거 등의 데이터 수집도 함께 진행. • 카메라는 한곳에 고정하고 움직이는 장면을 촬영. • 촬영할 여건이 된다면 저녁에도 촬영을 진행할 수 있으면 좋을 것임 • 자연객체를 따로 촬영하는 등 때에 따라 진행하도록 한다.		
촬영 분량	6시간		

2.2 규제관련 사항

- COVID-19의 상황을 주시하면서 현실적으로 촬영할 수 없는 곳의 협상보다는(e.g 박물관 등의 정부와 지자체 소유 건물, 혹은, 롯데월드 아이스링크 등 COVID-19 확진자 발생에 따라 방역 및 접근이 제한되는 장소 등) 새로운 로케이션을 발굴하면서 촬영 계획을 수립 및 지속적으로 변경함

2.3 획득 및 정제 절차

- 장면당 계획된 분량의 촬영을 하며, 수작업 또는 검출 도구를 통한 정제로 제거된 분량에 대한 즉각 추가 촬영을 진행하도록 함
- 수집된 촬영 데이터들은 촬영 현장에서 전문 촬영 프로듀서가 촬영 시나리오의 일치성 여부 검수, 연기 적절성 확인, 영상 상태 확인 과정을 확인

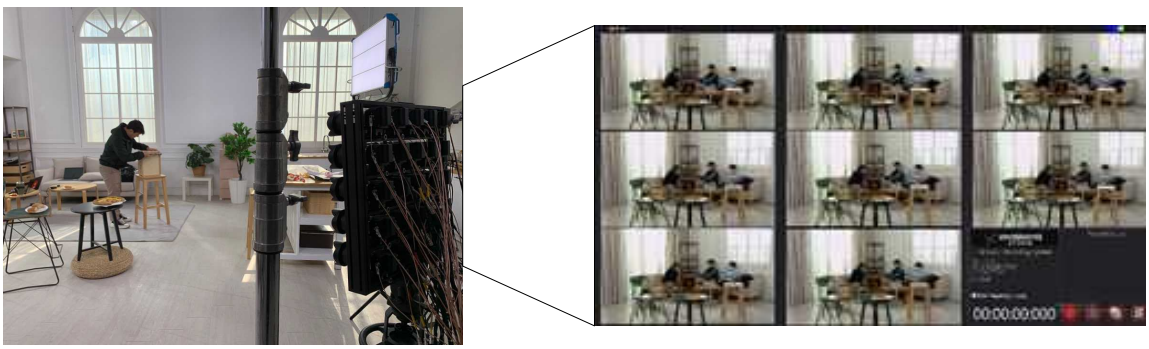
- 전문 프로듀서의 결정에 따라 편집 또는 재촬영을 진행하여 데이터 품질관리를 수행
- 데이터 삭제 과정에서 25개의 카메라로부터 촬영된 사진이 동시에 제거되도록 함
- 데이터 중복성 제거 : 수집을 통해 확보한 이미지를 대상체 별로 분류. 분류된 이미지를 솔루션을 통해 도출된 유사도에 따라 중복성이 확인된 이미지들을 제거
- 데이터 분류 작업 계획 : 기존의 LF dataset의 구조를 참고하여 디렉토리(Directory)를 구성하여 분류

① 카메라 캘리브레이션

- 실촬영 전에 5x5 카메라 정형, 컬러 캘리브레이션 진행.

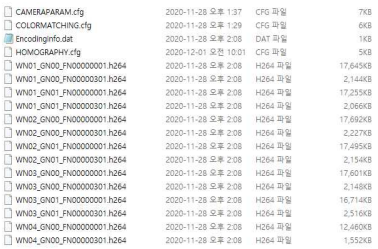
[그림 Ⅲ-82] 카메라 캘리브레이션

- 카메라 촬영



[그림 Ⅲ-83] 촬영 소프트웨어로 촬영 진행

② 데이터 정제 과정



H264 원본 파일

이미지 컨버팅

육안 및 Depth 추출로
캘리브레이션 품질 확인

비식별화 처리

데이터 가공을 위해
클라우드 업로드

[그림 Ⅲ-84] 데이터 정제 과정

2.4 획득 및 정제 기준

〈표 Ⅲ-67〉 데이터 정제 대상 예시

번호	정제 대상	예시
1	데이터 품질 이상	왜곡, 노이즈, 촬영 실수
2	데이터 중복성	같은 이미지 중복
3	개인정보 노출	방송 뉴스 이상의 얼굴 노출이나 차량 번호 노출 등
4	데이터 균형 분류	촬영 동작 및 객체 균형 확보를 위한 사전계획, 노출인원 포함
5	데이터 정량적 목표	총 75만장 이미지 확보. 촬영장소당 평균 5만장 목표.

3 어노테이션/라벨링

3.1 어노테이션 / 라벨링 절차

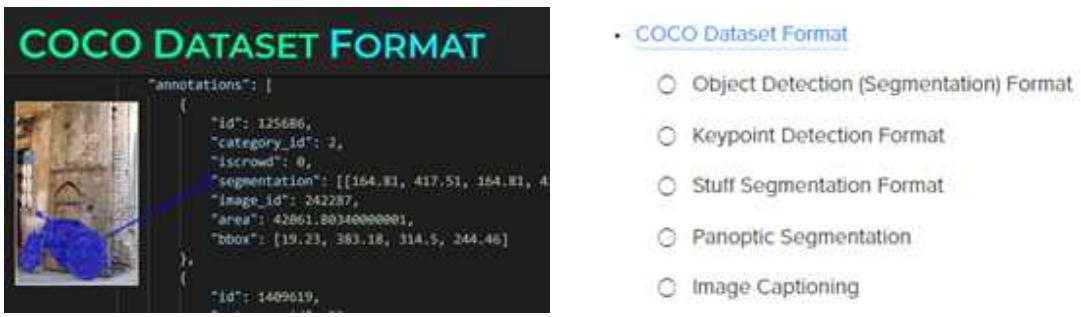
1) 고해상도 LF AI 학습용 데이터 어노테이션

- 카테고리 범주의 선정은 기존의 단어 분류체계를 사용하거나 COCO Dataset과 같은 기존 데이터 분류체계를 활용할 수 있음
- 또한, 응용 분야를 타게팅하여 Dataset을 구축하는 경우 직접 사용자 테스트를 거쳐 후보군을 수립할 수 있음
- COCO Dataset과 동일한 분류체계를 사용하는 경우 연구 활용에 있어 비교군이 될 수 있어 유효함
- 수집된 데이터에 대해 전체적인 데이터양에 비례하여 객체 최소 데이터양의 기준을 선정하여 Dataset의 균형을 맞춤

[그림 Ⅲ-85] 대표범주 Iconic Image 예시

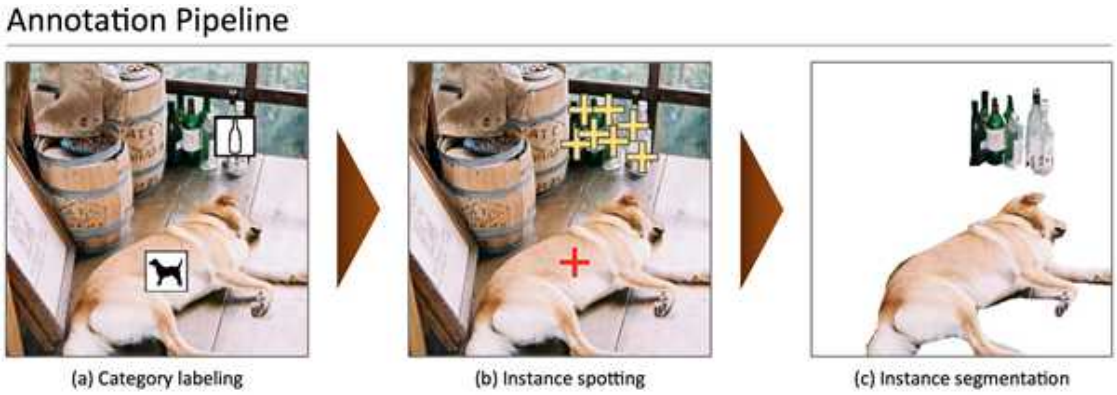
[그림 Ⅲ-86] 대표범주 Scene Image 예시

- 대표범주는 실체의 구분을 쉽게 지정할 수 있는 개체인 Iconic Image(e.g.동물, 사람, 의자)와 명확한 경계의 선정이 어려운 개체인 Scene Image(e.g.하늘, 숲, 바다)의 포함 여부를 선정을 우선함
- 세분화 범주는 계층구조의 세분화 정도를 의미하여 예를 들어 '동물'범주 하위 '개'로 명세하는 것 혹은 '치와와'로 더욱 세분화하여 명세하는 것을 의미함
- 전체범주와 부분범주는 예를 들어 사람과 얼굴같이 전체와 부분에 해당하는 범주를 의미함
- 응용 분야에서는 치와와를 추적 시에 강아지를 추적을 전처리하고, 얼굴 추적 시에 전체범주인 사람을 추적하여 탐색에 대한 전체 리소스를 줄일 수 있음



[그림 Ⅲ-87] Dataset Format 예시

- 다음 단계로 라벨링에 사용될 주석에 포함될 정보를 정의하여 Format 제작하는 과정을 진행함
- Annotation 정보는 기본적으로 색인을 위한 id, 카테고리 분류, 영역을 포함함. 각 정보는 레이블링 된 데이터의 카테고리별 검색 기능과 같이 응용프로그램에서 활용됨



[그림 Ⅲ-88] Annotation 파이프라인(COCO Dataset)

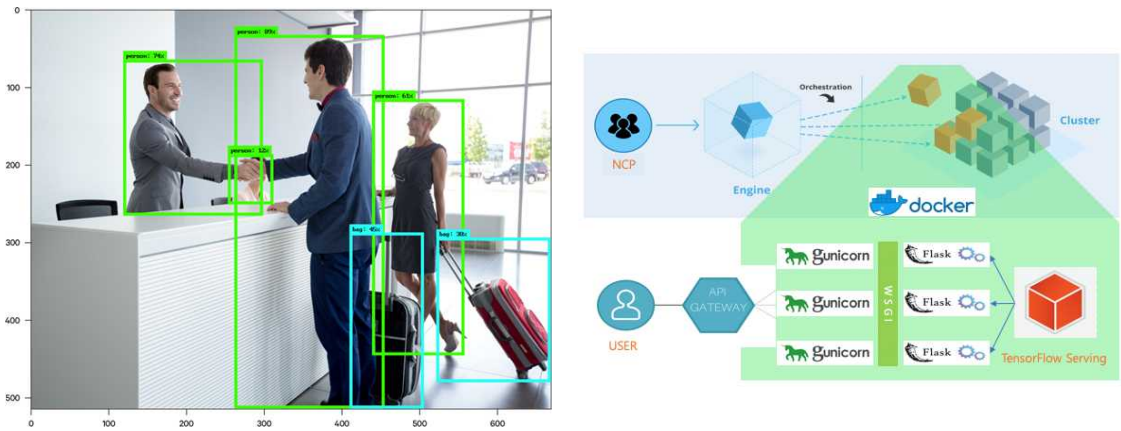
- 본 사업에서 제안한 고해상도 LF 데이터는 5 x 5의 카메라 배열로 촬영을 진행하여 단일카메라 촬영과 비교하여 25배의 데이터를 가지게 됨
- 객체 인스턴스에 라벨을 붙이는 비용 때문에 효율적이면서도 고품질 주석 파이프라인의 디자인이 중요하며 주석 파이프라인은 위 그림과 같이 3가지 과정으로 진행함

① 데이터 세트 주석에서는 가장 먼저 각 이미지에 어떤 객체 카테고리를 포함하는지 명시함. 이 과정에서 객체를 상·하위로 그룹화한 범주를 제공하여 라벨링 작업자의 작업량을 줄임

② 이미지의 객체 범주의 모든 각각의 인스턴스를 명시함

③ 최종적으로 각 객체 인스턴스를 세분화하는 과정으로 작업자가 지정한 객체 인스턴스를 분할함

2) 고해상도 LF AI 학습용 데이터 초기데이터 생성을 위한 자동화 방안



[그림 Ⅲ-89] 상용 ObjectDetection 서비스 예시 및 아키텍처(네이버 클라우드 플랫폼)

- 인스턴스 수가 큰 이미지의 경우에 작업자가 Annotation 의 정확도가 떨어지고 작업이 지연될 수 있어, 이를 방지하기 위한 초기 데이터 생성의 자동화 방안을 사용함. 이미지 내의 객체 (사람이나 차량과 같은) 식별은 네이버 클라우드 플랫폼의 Object Detection, 혹은 유사 기능을 제공하는 서비스를 도입하여 수행함
- 초기 식별된 객체 탐지의 결과로 객체의 이름, 바운딩 박스, 탐지정확률을 확인할 수 있으며 이는 Annotation 과정 중 Instance Segmentation 과정의 이미지 분할의 보조로 쓰이며, 이미지 내 오브젝트 크기가 작은 인스턴스에 대한 객체 인식 자동화에 활용함
- -COCO Dataset의 분류를 기준으로 하는 네이버 플랫폼의 Instance Segmentation 과정에서 Object Detection 후 교정작업을 진행하여 이미지 내에서 인스턴스를 분할함. 다만 데이터 분류가 다른 경우 이를 치환하는 모듈의 개발을 진행함
- 프로젝트 초기에 LF 데이터의 분석을 위한 자동화 도구로 2종 이상의 상용 툴, 오픈소스, 혹은 국책연구기관의 기술 등을 면밀히 검토하여 필요한 솔루션을 도입하거나 이와 연동하여 사용할 수 있는 도구를 개발하여 프로젝트를 수행함

3) 고해상도 LF AI 학습용 데이터 저작도구

- 카테고리의 라벨링 과정은 이전 과정에서 정의한 포맷을 기준으로 각 이미지가 어떤 객체 카테고리를 포함하고 있는지 명세하는 작업임
- 주석 작성은 이미지 기반 객체 감지를 위한 교육 데이터를 만들기 위해, 다양한 기능과 효율적으로 이미지 레이블을 지정하도록 설계된 이미지 주석 도구를 제작하여 진행함

[그림 Ⅲ-90] Annotation 도구 예시(Supervisely)

[그림 Ⅲ-91] Annotation 도구 예시 (COCO Annotator)

- 이미지 세그먼트에 레이블을 지정하고, 객체 인스턴스를 추적하고, 보이지 않는 부분이 있는 객체에 레이블을 지정하는 기능을 기본으로 함
- 이미지에서 객체 영역 정의는 수동으로 정의하고 텍스트 설명을 작성하며 마스킹 도구를 통해 직접 경계 상자로 표시하거나 포함 영역을 정의하기 위해 점을 표시하여 객체를 표시할 수 있게 하는 등 직관적으로 구성함
- 사용자가 Annotation을 직관적으로 편집할 수 있도록 인터페이스 제공하고 자유형 커브 또는 다각형을 사용하여 이미지에 주석을 다는 등의 사용자 편이 기능을 설계하여 저작도구를 개선함

3.2 어노테이션 / 라벨링 기준

- 카테고리 범주의 선정은 기존의 단어 분류체계를 사용하거나 COCO 데이터셋과 같은 기존데이터 분류체계를 활용할 수 있음
- 또한, 응용 분야를 타게팅하여 데이터셋을 구축하는 경우 직접 사용자 테스트를 거쳐 후보군을 수립할 수 있음
- COCO 데이터셋과 동일한 분류체계를 사용하는 경우 연구 활용에 있어 비교군이 될 수 있어 유효함
- 대표범주는 실체의 구분을 쉽게 지정할 수 있는 개체인 Iconic Image(e.g.동물, 사람, 의자)와 명확한 경계의 선정이 어려운 개체인 Scene Image(e.g.하늘, 숲, 바다)의 포함 여부를 선정을 우선함
- 세분화 범주는 계층구조의 세분화 정도를 의미하여 예를 들어 '동물'범주 하위 '개'로 명세하는 것 혹은 '치와와'로 더욱 세분화하여 명세하는 것을 의미함
- 전체범주와 부분범주는 예를 들어 사람과 얼굴같이 전체와 부분에 해당하는 범주를 의미함
- 응용 분야에서는 치와와를 추적할 시에 강아지를 추적을 전처리하고, 얼굴 추적 시에 전체범주인 사람을 추적하여 탐색에 대한 전체 리소스를 줄일 수 있음
- 다음 단계로 라벨링에 사용될 주석에 포함될 정보를 정의하여 Format 제작하는 과정 진행함
- Annotation 정보는 기본적으로 색인을 위한 id, 카테고리 분류, 영역을 포함한다. 각 정보는 레이블링 된 데이터의 카테고리별 검색 기능과 같이 응용프로그램에서 활용된다.

- 본 사업에서 제안한 고해상도 LF 데이터는 5 x 5의 카메라 배열로 촬영을 진행하여 단일카메라 촬영과 비교하여 25배의 데이터를 가지게 됨.
- 객체 인스턴스에 라벨을 붙이는 비용 때문에 효율적이면서도 고품질 주석 파이프라인의 디자인이 중요하며 주석 파이프라인은 위 그림과 같이 3가지 과정으로 진행함

① 데이터 세트 주석에서는 가장 먼저 각 이미지에 어떤 객체 카테고리를 포함하는지 명시함. 이 과정에서 객체를 상·하위로 그룹화한 범주를 제공하여 라벨링 작업자의 작업량을 줄임

② 이미지의 객체 범주의 모든 각각의 인스턴스를 명시함

③ 최종적으로 각 객체 인스턴스를 세분화하는 과정으로 작업자가 지정한 객체 인스턴스를 분할함

3.3 어노테이션 / 라벨링 교육

〈표 Ⅲ-68〉 어노테이션/라벨링 교육 내용 및 방법

교육 구분	내용	수행 기간	비고
기본 교육	• 카메라 켈리브레이션 이해 • 직접 영상 획득 방법과 장치에 대한 이해 • 딥러닝에 대한 이해 및 객체 인식 알고리즘에 대한 이해	-	온/오프라인

3.4 어노테이션 / 라벨링 도구 및 사용법

- Lightfield 카메라 획득 시스템
  - 25대 LF카메라와 같이 세트로 4대의 데스크탑 서버를 운영. 서버에서는 각 카메라의 동기화 프로그램이 작동
  - 동시 촬영 지원 소프트웨어 사용
  - 각 카메라의 프리뷰를 실시간으로 모니터 4대에서 확인
  - 10초마다 각 인코더에서 녹화된 h264 영상 파일을 저장
  - 하나의 인코더에서 4대의 카메라를 담당. 총 4대의 인코더에서 인코딩을 진행하며 데이터 저장

- 정제 도구
  - ImageRectification : 고해상도 LF 이미지 원시데이터에 카메라 정형, 색 보정을 적용하는 프로그램
  - ffmpeg : 디코딩 프로그램. h264=〉png 변환
  - python script : 프레임 단위로 이미지 시퀀스 정리, imageRecification, ffmpeg 실행
  - GA-NET : GA-NET 뉴럴네트워크를 활용한 고해상도 LF 데이터 Depth 추출 기능. 사전 캘리브레이션 품질 체크 방법 중 하나

[그림 Ⅲ-92] 고해상도 LightField 카메라 시스템 세트

4 데이터 검수

4.1 검수 절차

1) 현장에서의 1차 검수 진행
  - 고해상도 LF 데이터 수집 소프트웨어로 상시 촬영 프리뷰를 검토
  - 잘못된 촬영 영상, 소프트웨어 에러 등이 발견되면 해당 촬영분 기록하여 삭제
  - 촬영 전 단계에서 조명을 검토. 프리뷰 영상에 따라 조명셋팅 상황 수정
  - 일반인 통제. 부득이하게 촬영된 경우 해당 촬영분을 즉시 삭제
  - 캘리브레이션 사전 검토. 촬영 소프트웨어에서 캘리브레이션 결과를 프리뷰로 출력. 캘리브레이션에 문제가 되는 케이스는 Merge에서 제외 처리

2) 완료된 데이터를 데이터가공을 위한 데이터가공팀에서 2차 검수함

- 촬영분 검토. 촬영 객체, 동작 리스트 확인
- 캘리브레이션 적용. 5x5 기하 보정 결과 값 최종 확인. 문제 발생시에는 캘리브레이션 cfg 데이터 에러파일을 찾고 제외처리. 이후 다시 Merge하여 캘리브레이션 cfg데이터 생성
- 비식별화 대상 장면 검토. 문제사항 발견시 해당 씬은 삭제 처리, 혹은 blurring 처리

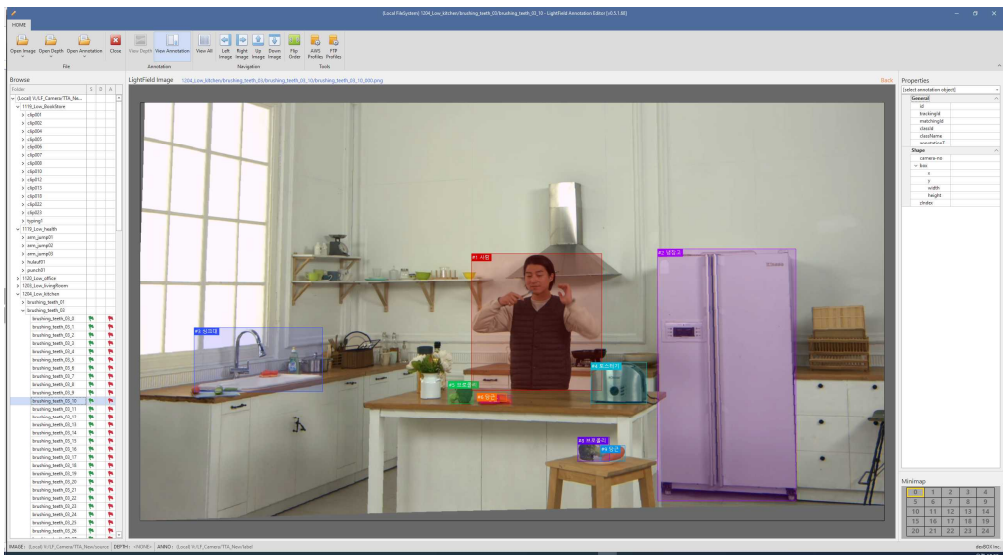
3) 2차 검수된 데이터를 통해 가공된 데이터를 4인 이상으로 편성된 외부 데이터 검수팀이 월간 1회(혹은 총 3회)의 외부 데이터 검수를 수행함

4) 외부 데이터 검수팀 구성은 NIA가 보유하고 있는 데이터 검수 기 경험자 혹은 기관을 선정거나 SuperAI 등의 기존 데이터 구축 및 검증 경험을 보유한 업체 혹은 대학과 협력도록 하겠음

5) 최종적으로 TTA로부터 NIA의 타 과제에서의 TTA 품질평가 가이드라인을 토대로 데이터를 납품함

※ 검수/저작 도구

- LightField Annotation Editor
- 참여기관 데브박스에서 제작한 고해상도 Lightfield 이미지 AI 데이터 전용 저작도구
- TTA 초기 검수용 저작도구로써 설치파일 업로드 완료



[그림 Ⅲ-93] 저작도구로 라벨링 작업결과 시각적 확인 장면

4.2 검수 기준

〈표 Ⅲ-69〉 비식별화 검수 기준

구분	작성 지침
ID	1-1
항목명	비식별화 검사
내용	초상권, 개인정보 등을 침해할 소지가 있는 촬영 내용을 확인. 문제가 되는 분량은 삭제처리
지표	비식별화 대상 이미지
목표(기준)	최대 750,000개
검증 환경	촬영현장, 일반 PC 환경
검증 절차	• LF 카메라로 촬영한 이미지 파일을 육안으로 확인 • 초상권, 개인정보 침해 요소를 확인하고 문제되는 촬영분은 삭제 처리 및 삭제 내용 기록 • 삭제하지 않고 데이터 확보해야하는 경우 문제되는 범위에 blurring처리를 한다
결과	비식별화 검수에서 모두 통과한 수집데이터 삭제 처리 혹은 blurring처리 이력

〈표 Ⅲ-70〉 카메라 캘리브레이션 검수 기준

구분	작성 지침
ID	2-1
항목명	5 x 5 LF 카메라 캘리브레이션
내용	촬영단계에서 캘리브레이션으로 수집한 cfg데이터로 수집 원시데이터에 적용한 결과를 검수
지표	1. 체스보드 코너점 영상점 전차 2. 실내공간에서 객체점 분해력 두가지 명시
목표(기준)	1. ±1 pixel 이내의 정확도 2. ±3.2cm 이내의 정확도 명시
검증 환경	외부 기관(동서대학교) 검증
검증 절차	캡쳐한 체스보드 이미지와 계단형 입체 마크를 제작하여 제3자 전문가 평가 자체 LF 카메라 캘리브레이터 결과를 대상으로 국내 영상처리 전문가인 이병국 교수 (동서대학교)의 엠비언트인 텔리전스 연구소가 카메라 간, 객체별 영상점 잔차를 확인
결과	칼리브레이션 된 결과

1) nxn Camera(n=5,3,1) 내부/외부 파라미터(internal/external parameter) 획득에 관한 전문가 검증 - 일반 연구실 실내 공간(6mx6m)에서 영상 해상도 1280x720 기준에서 영상점 잔차는 ±1 pixel 이내, 객체점은 약±3.2cm 이내의 정확도로 획득함

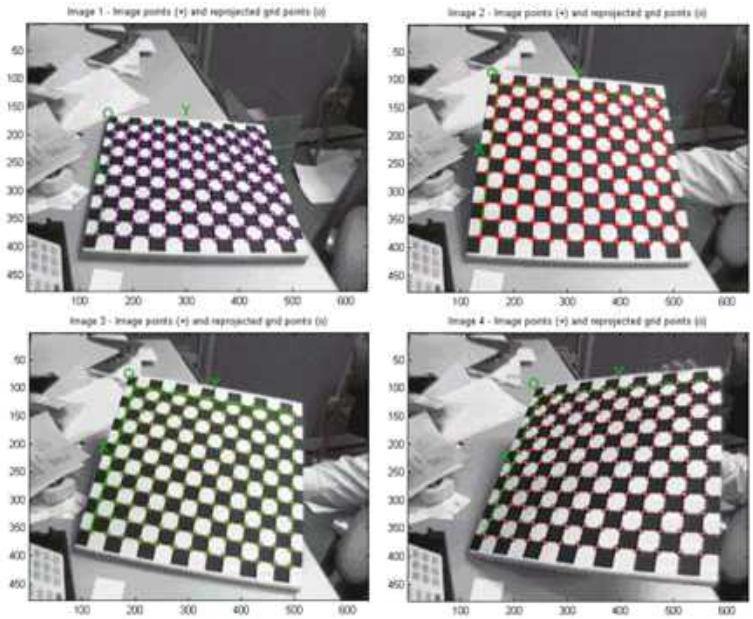
- 체스 보드(chess board) 캘리브레이터(calibrator)를 다양한 각도에서 캡처하였는지 확인

- 체스 보드 이미지에서 코너점들이 정확하게 잘 찾아졌는지 확인
- 계산된 내부/외부 파라미터를 이용하여 별도의 기준 계단형 입체 마크를 제작하여 정확도 확인

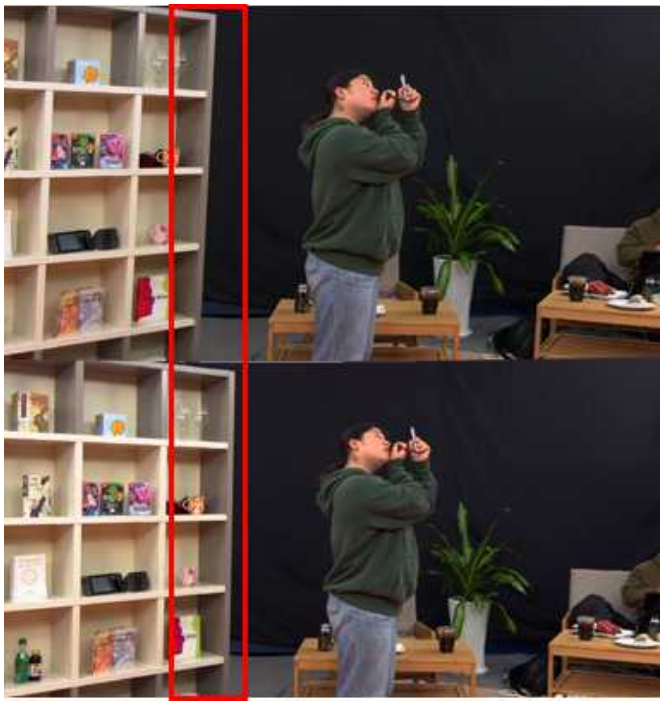
2) nxn Camera(n=5,3) 외부 파라미터(external parameter)는 현장 촬영전 카메라 구조물 이동에 따른 카메라 간의 이동을 염려하여 필요에 따라 현장용 켈리브레이터(calibrator)를 제작하여 실내와 같은 정확도 확보

3) 획득한 영상 프레임별 기존의 딥러닝(Deep Learning) 객체 인식(Object Detection) 알고리즘으로 기본 Bounding Box 및 라벨(label)을 구하고 오퍼레이터(operator)들이 수정하여 정확도 확보

4) 오퍼레이터(operator)들에게 카메라 켈리브레이션(camera calibration)의 의미와 방법, Plenoptic Camera에 대한 이해, 집적 영상(integral image)의 획득 방법과 장치에 대한 이해, 딥러닝(Deep Learning)에 대한 이해, 객체 인식(Object Detection) 알고리즘에 대한 이해 등 세미나와 교육을 통하여 데이터 품질관리 확보
   - 위탁기관 동서대학교 이병국 교수 IAI 연구소와의 연구 협력을 통하여 품질검증 협력방안 마련
   - 과제 관련 오퍼레이터들에게 online/offline 교육을 통하여 데이터 품질관리 확보
   - 카메라 켈리브레이션관련 체스보드 캘리브레이터, 현장용 캘리브레이터 그리고 계단형 입체 마크를 제작



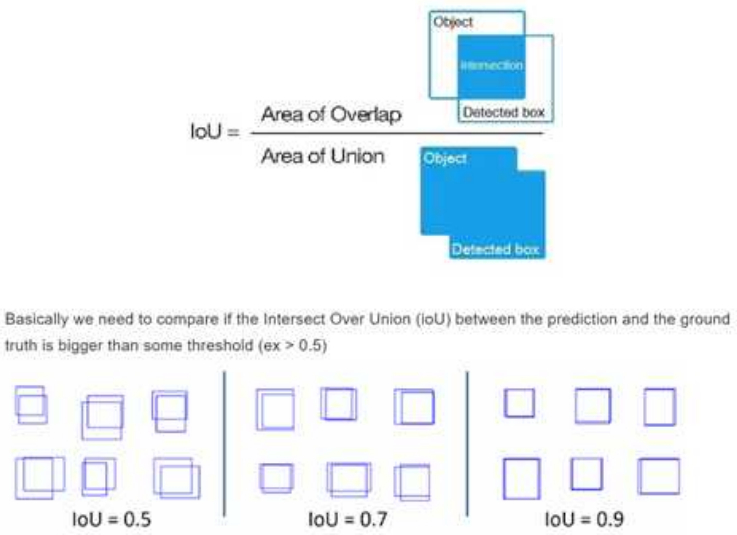
[그림 Ⅲ-94] nxn 체스보드를 활용한 카메라 캘리브레이션



[그림 Ⅲ-95] 잘못된 보정의 예(위), 정상 보정의 예(아래)

〈표 Ⅲ-71〉 바운딩 박스 정확도 검수 기준

구분	작성 지침
ID	3-1
항목명	바운딩 박스 정확도
내용	구축가이드 라인의 어노테이션 구조와 실제 구축된 어노테이션 데이터의 구조를 비교하여 내용을 확인
지표	IoU (Intersection Over Union)
목표(기준)	0.75 이상
검증 환경	일반 PC 환경
검증 절차	• 검증된 라벨러가 작성한 결과를 GT(Ground Truth)로 정의함 • 기구축된 데이터의 어노테이션 결과의 넓이와 GT를 비교, 두 결과가 얼마나 일치하는지를 확인 • 검수를 통과하지 못한 경우 워커에게 라벨링 작업 수정 요청
결과	검증된 어노테이션 데이터



[그림 Ⅲ-96] Bounding Box의 정확도를 위한 측도(measure) IoU 계산식과 값의 의미

5 데이터 활용 방안

5.1 학습 모델

- **(학습 모델 개발 계획)**
  - 2D 영상 객체 분리 알고리즘으로 주로 활용되는 U-Net 또는 DeepLab v3++의 기본 네트워크 구조를 사용하되, 고해상도 LF 입출력의 크기 조정 및 중합 처리를 위한 CNN 레이어를 기본 네트워크의 입출력에 추가하여 테스트할 계획임
  - 부족한 훈련자료에서도 높은 성능을 발휘할 수 있는 모델을 학습하기 위해 ImageNet 등의 자료에서 pre-trained 모델을 fine tuning 하여 사용하는 전이학습(transfer learning) 기법을 활용할 계획임.
  - 효율적인 하이퍼 파라미터 탐색을 위해 AutoML 기술 중 하나인 베이지안 최적화(Bayesian Optimization) 기술 활용하고, 본 최적화의 Surrogate 모델로는 TPE(Tree of Parzen Estimators)를 이용함
  - AutoML 중 베이지안 최적화를 활용한 하이퍼파라메터 탐색은 hyperopt 오픈소스 라이브러리를 사용할 계획임 (http://hyperopt.github.io/hyperopt/)
  - 베이지안 최적화의 각 반복에서 안정적인 검증 성능 측정을 위해 k겹 교차검증(k-fold cross validation) 기법을 적용하고, 성능 지표로는 IoU(Intersection over Union)을 사용함
  - 베이지안 최적화를 활용하여 얻어진 최종 모델들(U-Net & DeepLab v3++)과 Soft Voting 기반의 Ensemble 모델의 성능을 비교하여 최종 모델 선정함
  - 동작인식 서비스 사용 신경망 구조로는 3D CNN과 2D CNN의 결합으로 동작을 인식하는 YOWO(You Only Watch Once)를 적용

5.2 서비스 활용 시나리오

- 주관기관의 4D Interactive 입체영상 제작의 자료처리 과정에 활용함으로써, 비용 절감 및 효율성 증대를 통해 사업화 및 부가가치 창출에 기여
- 고해상도 LF 카메라 자료처리에 대한 딥러닝 기술 적용성 검토 결과와 더불어 응용서비스 개발 과정에서 확인된 기술 이슈 공유를 통해, 장기적 관점에서 고해상도 LF 카메라 기술의 응용 분야 확대와 상용화를 통한 부가가치 창출에 기여

- 일반 이미지나 영상 자료와는 달리 고해상도 LF 카메라 자료는 개별 카메라의 독립적 영상 자료뿐만 아니라 상호 연관된 인접 정보를 포함하고 있음. 따라서 고해상도 LF 카메라 자료의 인접 특성을 고려한 딥러닝 모델을 개발함으로써 일반 영상 자료를 이용한 객체 분리 알고리즘보다 효율성이 높고 품질이 향상된 고품질의 객체 분리 결과를 획득할 수 있음.
- 이러한 동작 인식 결과에 따라, 동작의 주체를 분리(Localization)함으로써, 고해상도 LF 영상 자료를 라벨링 하거나, 사용자의 다양한 목적에 따라 가상환경 등의 영상합성 등과 같은 후처리 공정에 활용할 수 있음.
- 고해상도 LF 카메라의 대용량 자료에 대한 초점 변경, 디포커싱 등 영상처리를 수행하는 데 있어서 하드웨어 자원이 많이 필요한 이슈를 딥러닝을 통한 동작 인식을 이용하여 해결하고 효율성과 처리 시간 단축 등의 향상이 기대
- 향 후 현재 대부분 수작업으로 이뤄지고 있는 객체/배경 분리(로토스코핑)을 자동으로 정밀하게 분리하기 위한 방향으로 활용 가능
- 즉, 객체의 동작을 인식하여 목표하는 객체를 인지하고, 해당 객체와 나머지 배경을 분리하는 2단계 분리 방법을 적용함으로써 보다 정밀한 분리가 이뤄짐

제8장 폐암 AI 학습 데이터

1 데이터 정보 요약

1.1 가이드 분류

대분류	이미지	중분류	헬스케어	소분류	X-ray, CT, PET/CT

1.2 데이터 정보

데이터 이름	폐암 AI 학습 데이터
활용 분야	의료
데이터 요약	데이터 유형별(X-ray, CT, PET/CT) 양성/악성/정상 구축

1.3 데이터 구축 개요

1.4 구축 목적

- 인공지능 기반 흉곽 구조를 이해하는 단순 흉부영상 판독 기술 정확도 향상
- 인공지능 기반의 흉부 전산 단층촬영의 영상판독 기술 정확도 향상 및 폐암 예측도 향상
- 간유리음영 소결절에 대한 검사자 간의 판독 결과의 차이와 검사 시간 최소화 달성
- PET/CT의 소결절에 대한 폐암예측도를 향상하여 최종 조직검사와 비교하여 검사의 정확도 판단에 보조기능 제공

1.5 활용 분야

- 단순 흉부영상, 흉부 전산단층촬영 영상, PET/CT 영상 판독 기술을 종합적으로 판단하여 최종적으로 악성의 예측도를 제고하는 프로그램 개발 및 실증 데이터로 활용

- 폐암의 진단 및 예후 예측을 위한 인공지공 기반 예측 요인(feature) 개발을 위한 실증 사업 솔루션 개발 및 사업화 확장

1.6 유의 사항

- 환자의 의료정보가 포함되어 있는 의료 데이터(원시 데이터)는 승인된 연구자만 접근 가능
- 페보건복지부의 보건의료 데이터 활용 가이드라인에 따라, 데이터 활용 및 제 3자의 배포를 위해서는 해당 의료 기관의 데이터 심의기관의 허가를 받아야 함

2 데이터 획득 및 정제

2.1 원시 데이터 선정

〈표 Ⅲ-72〉 원시데이터 유형 및 목표 건수

유형	목표 건수(건)		
	분류	환자 수	영상 수
X-ray	양성	500명	500장
	악성	3,000명	3,000장
	정상	10,000명	10,000장
흉부 CT	양성	1,000명	150,000장
	악성	2,500명	375,000장
	정상	1,000명	150,000장
PET/CT	양성	500명	125,000장
	악성	3,000명	750,000장
	정상	1,000명	250,000장

2.2 규제관련 사항

- 원시 데이터 획득을 위해 고신대학교 복음병원의 IRB 승인 획득
- 환자 비식별화 및 익명화 작업을 수행하여 의료 관련 개인정보보호법에 침해요소 제거

2.3 획득 및 정제 절차

〈표 Ⅲ-73〉 데이터 획득 및 정제 절차

구분	내용	
데이터 획득	고신대학교 복음병원의 의료데이터베이스(EMR, PACS)에서 환자의 폐암 관련 X-ray, CT, PET/CT 영상을 획득	
데이터 정제	공통 사항	• 익명화 SW를 이용해 환자 ID, 영상 ID를 수집 데이터베이스에 의해 관리되는 기준 ID로 치환하고, 임상 데이터와 함께 가공 기관 및 관련 정보가 함께 관리될 수 있도록 데이터베이스를 구축 • Dicom 헤더에 포함된 환자 정보를 제거 • 영상 내에서 Tagging 혹은 기기별 영상 레이아웃에 의해 포함되는 환자 식별 정보를 제거

2.4 획득 및 정제 기준

〈표 Ⅲ-74〉 데이터 획득 및 정제 기준

유형	수집기준	배제기준
X-ray	• Chest PA 영상만을 사용 • 정상/양성/악성 기준 - 정상: CT에서 정상으로 확인된 케이스 - 양성: 조직학적으로 양성으로 확인되거나 CT에서 경과 관찰 중 변화가 없는 경우 - 악성: 생검 및 수술로 악성으로 확진된 병변으로, CT 판독에서 nodule이 0.4cm이상인 환자의 X-ray 영상	• 뚜렷한 인공음영이 있는 경우, Chest AP영상은 제외 • CT에서 확인된 결절 중 영상의학과 전문의의 판단에 결절이 보이지 않는 경우는 수집 대상에서 제외함
CT	• 정상/양성/악성 기준 - 정상: CT에서 정상으로 확인된 케이스 - 양성: 조직학적으로 양성이 확인되거나 흉부 CT에서 경과 관찰 중 변화가 없는 경우 - 악성: nodule이 0.4cm이상인 환자의 X-ray 영상 - Slice Thickness: 2~10mm 사이의 영상을 대상으로 함	• 움직임에 대한 인공음영이 있는 경우 • Slice spacing이 있는 경우는(절편 사이의 간격)영상 수집에서 제외함
PET-CT	• 정상/양성/악성 기준 - 정상: 폐와 종격동 임파선에 PET negative를 보인 경우 - 양성: 폐에 결절 가진 환자 중 조직학적으로 양성으로 확인되거나 2년 경과 확인 시 변화가 없는 경우 - 악성: 생검 또는 수술로 악성이 확인된 경우	• DICOM아닌 Image 형태의 영상

3 어노테이션/라벨링

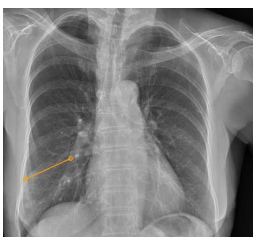
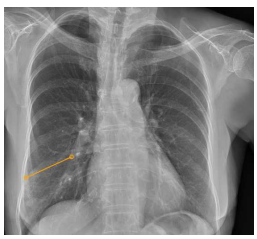
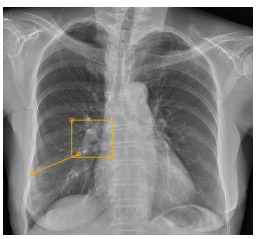
3.1 어노테이션 / 라벨링 절차

〈표 Ⅲ-75〉 어노테이션/라벨링 절차

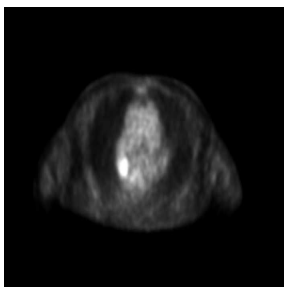
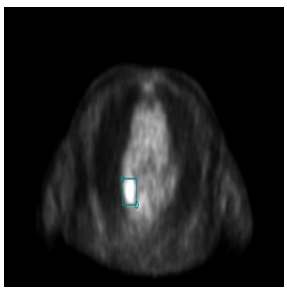
구분	X-ray	CT	PET/CT
1차 수행	• 수행인: 영상의학과, 호흡기내과 전문의 • 수행 내용: 병변부위 라인 체크	• 수행인: 영상의학과, 호흡기내과 전문의 • 수행 내용: 병변부위 라인 체크	• 수행인: 핵의학과,호흡기내과 전문의 • 수행 내용: 병변부위 바운딩 박스 체크
2차 수행	• 수행인: 별도 교육을 받은 인력 • 수행 내용: 병변부위 바운딩 박스 체크	• 수행인: 별도 교육을 받은 인력 • 수행 내용: 병변부위 폴리곤 체크	해당사항 없음

3.2 어노테이션 / 라벨링 기준

〈표 Ⅲ-76〉 X-ray 어노테이션/라벨링 기준

<table>
<tr><th>유형</th><th colspan="2">어노테이션 항목</th><th>어노테이션 수행인</th></tr>
<tr><td rowspan="4">X-ray</td><td colspan="2">• 1차 작업
– 병변 부위 라인 체크
(처리 데이터 수가 많아, 1차 작업(프리어노테이션)을 전문의가 수행한 후 교육을 수행한 인력이 2차 작업을 수행함)</td><td rowspan="2">• 전문의
– 영상의학과
– 내과 전문의</td></tr>
<tr><td>[수행 전]
  </td><td>[수행 후]
  </td></tr>
<tr><td colspan="2">• 2차 작업
– 병변부위 Bounding box 표시</td><td rowspan="2">• 별도의 교육을 받은 교육생</td></tr>
<tr><td>[수행 전]
  </td><td>[수행 후]
  </td></tr>
</table>

〈표 Ⅲ-77〉 CT 및 PET/CT 어노테이션/라벨링 기준

<table>
<tr><th>유형</th><th colspan="2">어노테이션 항목</th><th>어노테이션 수행인</th></tr>
<tr><td rowspan="4">CT</td><td colspan="2">• 1차 작업
– 병변부위 라인 체크</td><td rowspan="2">• 전문의
– 영상의학과
– 내과 전문의</td></tr>
<tr><td>[수행 전]
  </td><td>[수행 후]
  </td></tr>
<tr><td colspan="2">• 2차 작업
– 병변부위 폴리곤 표시</td><td rowspan="2">• 데이터 레이블링 관련 교육을 수료한 의공학, 공학, 의학 등 전공자</td></tr>
<tr><td>[수행 전]
  </td><td>[수행 후]
  </td></tr>
<tr><td rowspan="3">PET/CT</td><td colspan="2">◦병변부위 바운딩 박스 체크</td><td rowspan="3">• 전문의
– 핵의학과 전문의
– 내과 전문의</td></tr>
<tr><td>[수행 전]
  </td><td>[수행 후]
  </td></tr>
</table>

〈표 Ⅲ-78〉 어노테이션/라벨링 작업자 교육

교육 구분	내용	수행 기간	비고
기본 교육	• 암종의 이해 • 데이터 형태에 대한 이해 • 어노테이션 도구 교육 • 실습	• 암종별 1일 • 반복 수행	온/오프라인
심화 교육	• 암종의 세부 교육 • 데이터 어노테이션 방법 • 데이터 어노테이션 결과 평가 방법 • 실습	• 암종별 1주 • 반복 수행	온 /오프라인

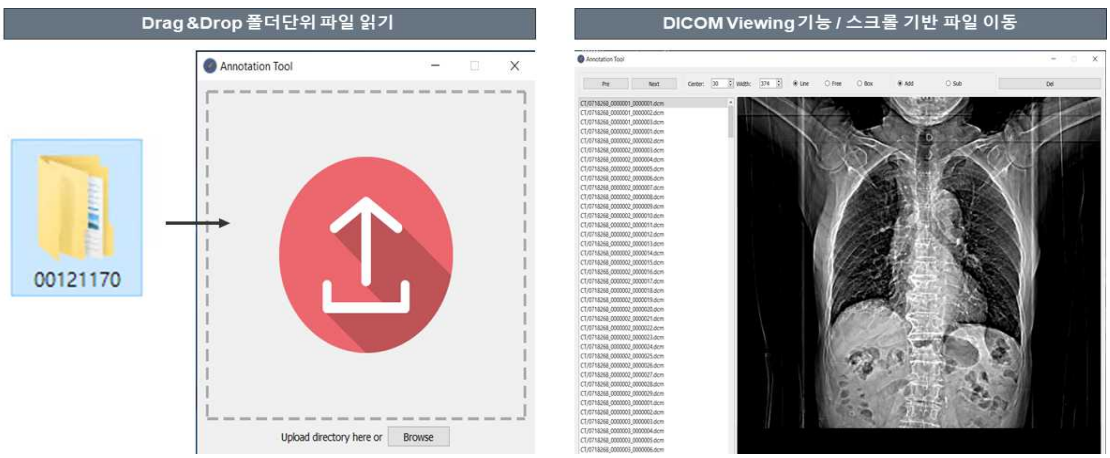
3.4 어노테이션 / 라벨링 도구 및 사용법

1) 저작도구

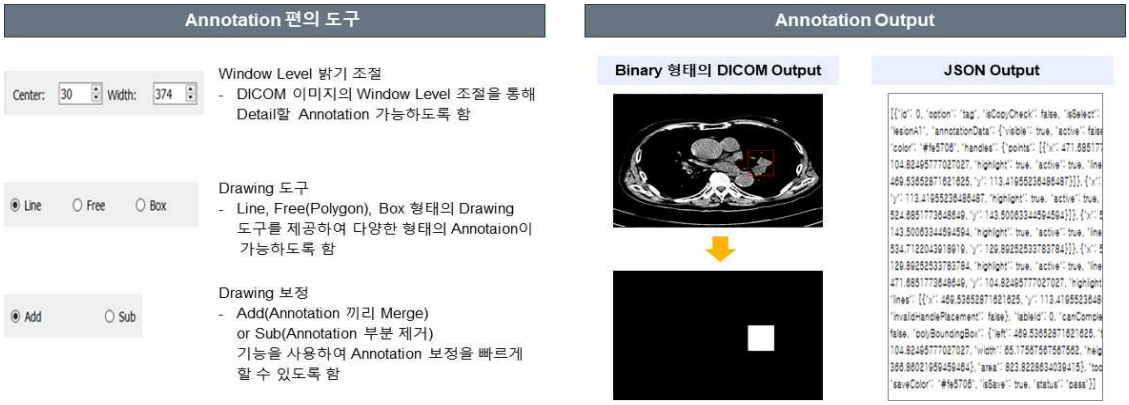
- 학습용 데이터 저작도구는 양성 영역 경계 어노테이션, 종양영역 경계 어노테이션, 자동 어노테이션 결과 저장 등을 제공하는 경북대학교 산학협력단에서 개발한 Window/Python 기반 프로그램을 사용함

2) 기능

- 폴더 읽기 / DICOM VIEWING 기능
- Annoation Drawing 편의 및 Ouput

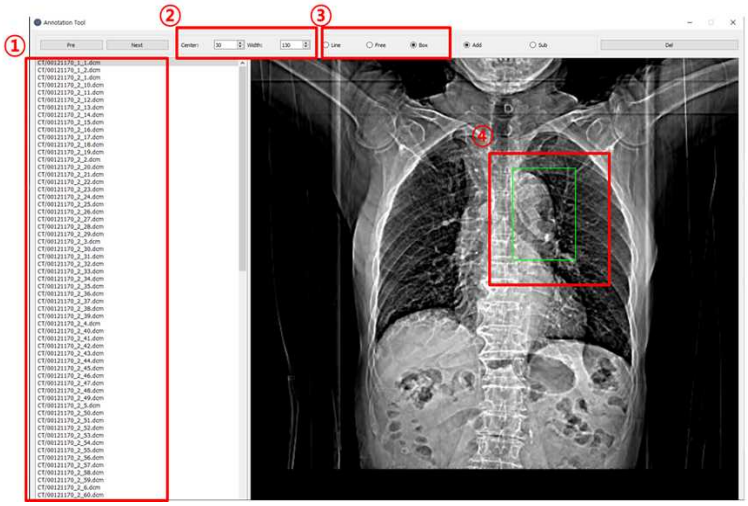


[그림 Ⅲ-97] 학습용 데이터 저작도구 폴더 읽기 / DICOM VIEWING 기능



[그림 Ⅲ-98] 학습용 데이터 저작도구 Annoation Drawing 편의 및 Ouput 기능

3) 절차



① 어노테이션 할 이미지 선택
(마우스 스크롤로 이동 가능)

② 필요 시 이미지 window level 조정

③ 어노테이션 종류 선택
(처음에 한번 선택 시 변경 전까지 선택이 유지됨)

④ 어노테이션 수행

[그림 Ⅲ-99] 어노테이션 작업 화면 예시

4 데이터 검수

4.1 검수 절차

1) X-ray 및 흉부 CT 데이터 어노테이션 및 검수 방법

- 1단계: 전문의가 1차 프리어노테이션 수행
- 2단계: 교육생이 2차 어노테이션 진행
- 3단계: (1차 검수) 파일 생성 여부 확인(어노테이션 된 데이터(Binary DICOM)와 메타데이터(JSON)파일이 올바르게 생성되었는지 파이썬 스크립트로 검수 진행
- 4단계: (2차 검수) 어노테이션 정확성 검토: 영상의학과 전문의에 의해 교차검증 진행

2) PET/CT 데이터 검수 방법

- 핵의학과, 내과 전문의가 병변 부위 bounding box 표시

4.2 검수 기준

1) 데이터 획득 단계(원본추출)

- PACS에서 영상추출을 위한 환자 대상자선정 및 추출 리스트 검토 2인 이상 확인 (암종별 정보, 환자정보 누락은 없는지, 정확한 검사 종류,날짜,시간 등)
- PACS에서 익명화하여 추출한 영상 파일을 원내 DATA서버로 옮겨 데이터 적절성 교차 확인 (DICOM head 정보, 영상의 누락은 없는지, 매핑정보(new ID)
- EMR ID 와 NEW ID의 매핑정보 재확인 후 클라우드 서버에 업로드

2) 2차 검수

- 교육생이 어노테이션을 시행하는 경우 90%의 정확도를 기준으로 90% 이상 일치 시 통과, 정확도 90% 미만 시 annotation 재수행
- 2차 검수 시 2명의 영상의학과 의사가 교차 검증(의견이 다를 경우 최종 검수자가 병리 결과 및 EMR, PACS 데이터를 확인하여 최종 결정함)

5 데이터 활용 방안

5.1 학습 모델

- 학습 모델 선정 근거
  - 이미지에서 객체를 인식하고 분류하는 다양한 알고리즘 중에서 암 병변 여부를 판단하는 딥러닝 알고리즘으로 DeepLab v3+를 권고함
  - Semantic Segmentation 은 컴퓨터비젼 분야에서 가장 핵심적인 분야 중 하나이며 이미지 내에 있는 물체들을 의미 있는 단위로 분할
  - Semantic Segmentation은 Deep Convolution Neural Network (깊은 신경망)을 적용해서 많은 발전 진행
  - DeepLab은 semantic segmentaion을 잘 해결하기 위한 방법으로 atrous convolution을 적극적으로 활용
  - 일반적인 convolution은 feature가 sparse하게 추출된 반면, atrous convolution의 경우 feature가 더 두드러지게 추출 가능
- 학습 모델 개발 계획

〈표 Ⅲ-79〉 폐암 AI 학습용 데이터 인공지능 학습 모델 개발 계획

유형	모델 개발 계획	학습 분배량
X-ray	• 인공지능 모델 개발 목표 – 병변 유무 판단 – 병변 위치 표기 – 병변 양성/악성 진단 • AI 베이스 라인 모델 – 수집 데이터에서 성능 향상을 위한 모델 개선, 신규 모델 적용을 수행하며, 목표 성능을 상회하는 모델에 대한 소스코드 등을 공개함	* 학습 : 시험 : 검증 비율은 환자수를 기준으로 8:1:1로 구성되며, 이미지 개수는 차이가 발생할 수 있음.
CT	• 인공지능 모델 개발 목표 – 병변 유무 판단 – 병변 위치 표기 – 병변 양성/악성 진단 • AI 베이스 라인 모델 – 딥러닝 알고리즘을 적용한 이미지 분석을 통해 각종 암 병변 여부를 판단하는 딥러닝 모델들로 DeepLab, VGG, InceptionNet과 같은 모델을 사용	* 학습 : 시험 : 검증 비율은 환자수를 기준으로 8:1:1로 구성되며, 이미지 개수는 차이가 발생할 수 있음.

유형	모델 개발 계획	학습 분배량
	- 수집 데이터에서 성능 향상을 위한 모델 개선, 신규 모델 적용을 수행하며, 목표 성능을 상회하는 모델에 대한 소스코드 등을 공개함	
PET/CT	• 인공지능 모델 개발 목표 - 딥러닝 알고리즘을 적용한 이미지 분석을 통해 각종 암 병변 여부를 판단하는 딥러닝 모델들로 DeepLab, VGG, InceptionNet과 같은 모델을 사용 - 수집 데이터에서 성능 향상을 위한 모델 개선, 신규 모델 적용을 수행하며, 목표 성능을 상회하는 모델에 대한 소스코드 등을 공개함	* 학습 : 시험 : 검증 비율은 환자수를 기준으로 8:1:1로 구성되며, 이미지 개수는 차이가 발생할 수 있음.

5.2 서비스 활용 시나리오

〈표 Ⅲ-80〉 폐암 AI 학습용 데이터 서비스 활용 시나리오

과제명	데이터 유형	응용서비스
폐암 AI 학습용 이미지 데이터 구축	X-ray	• 병변 검출 및 탐지 • 병변의 종류(양성/악성)구분
	흉부 CT	• 병변 검출 및 탐지 • 병변의 종류(양성/악성) 구분
	PET/CT	• 병변 검출 및 탐지 • 폐 결절과 임파선의 양성/악성 감별 • 악성 폐 결절과 악성 임파선 검출

제9장 갑상선암 AI 학습 데이터

1 데이터 정보 요약

1.1 가이드 분류

대분류	이미지	중분류	헬스케어	소분류	X-ray, CT, PET/CT

1.2 데이터 정보

데이터 이름	갑상선암 AI 학습 데이터
활용 분야	의료
데이터 요약	데이터 유형별 양성/악성(초음파, 병리), 전이(Neck CT) 데이터 구축

1.3 데이터 구축 개요

〈표 Ⅲ-81〉 단계별 데이터 구축 개요

단계	과정	`내용
1단계	기초데이터 획득 및 정제	• 고신대학교 복음병원의 의료데이터베이스(EMR, PACS)에서 실제 환자 데이터 추출 • 갑상선암 3종 이미지 - 초음파 - Neck CT - 병리 이미지(세침흡인검사)
2단계	익명화 및 비식별화	자동 익명화 및 비식별화 SW를 개발하여 공통 규격의 데이터로 가공

단계	과정	내용		
3단계	데이터 어노테이션	초음파 이미지	• 어노테이션 항목 - 병변부위 바운딩 박스 체크	• 전문가 수행 - 내과, 외과(내분비) - 영상의학과 - 이비인후과 의사
		Neck CT	• 어노테이션 항목 1차 작업 - 병변부위 라인 체크 2차 작업 - 병변부위 바운딩 박스 체크	• 어노테이션 수행인 1차 작업 - 영상의학과 전문의 2차 작업 - 별도 교육을 받은 인력
		병리 이미지	• 어노테이션 항목 - 이상 부위 폴리곤 형태 체크	• 어노테이션 수행인 - 병리과 전문의
4단계	검증	• 데이터 생성 시 최소 2인의 크로스체크를 통한 품질관리 수행 • 갑상선 초음파, Neck CT: 영상의학과 전문의로 구성된 검증 인력을 구성하여 데이터 검수 • 병리 이미지: 전문의 어노테이션 후 1인 병리 전문의 데이터 검수 • 외부기관 TTA를 통한 데이터 품질 검증		
5단계	AI모델개발	• 병변 유/무 진단, 병변 검출 모델 개발 • 모델 성능 확인 및 제시		
6단계	응용서비스개발	• 이미지를 input하면 병변의 양성/악성 유무와 위치를 표시하는 프로그램 개발		

1.4 구축 목적

- 갑상선암 관련 초음파, CT , 병리 판독 기술의 정확도 향상
- 검사자간의 판독 결과의 차이와 검사 시간을 줄이기 위한 목적
- 최종 조직검사와 비교하여 검사의 정확도 판단

1.5 활용 분야

- 초음파, CT, 병리 판독 기술을 종합적으로 판단하여 최종 조직검사 예측
- 갑상선 유두암 감별 및 진단

1.6 유의 사항

- 환자의 의료정보가 포함되어 있는 의료 데이터(원시 데이터)는 승인된 연구자 이외에는 접근이 불가능함

- 보건복지부의 보건의료 데이터 활용 가이드라인에 따라, 데이터 활용 및 제 3자의 배포를 위해서는 해당 의료 기관의 데이터 심의기관의 허가를 받아야 함

2 데이터 획득 및 정제

2.1 원시 데이터 선정

〈표 Ⅲ-82〉 원시데이터 유형 및 목표 건수

유형	목표 건수(건)		
	분류	환자 수	영상 수
초음파	양성	1,040명	25,790장
	악성	3,960명	116,128장
Neck CT	전이있음	232명	45,143장
	전이없음	270명	56,030장
병리 이미지	양성	1,585명	2,731장
	악성	1,190명	2,111장

2.2 규제관련 사항

- 원시 데이터 획득을 위해 고신대학교 복음병원의 IRB 승인 획득
- 환자 비식별화 및 익명화 작업을 수행하여 의료 관련 개인정보보호법에 침해되지 않도록 처리

2.3 획득 및 정제 절차

〈표 Ⅲ-83〉 데이터 획득 및 정제 절차

구분	내용	
데이터 획득	• 고신대학교 복음병원의 의료데이터베이스(EMR, PACS)에서 환자의 갑상선 초음파, Neck CT 영상, 병리 이미지를 획득	
데이터 정제	공통 사항	• 익명화 SW를 이용해 환자 ID, 영상 ID를 수집 데이터베이스에 의해 관리되는 기준 ID로 치환하고, 임상 데이터와 함께 가공 기관 및 관련 정보가 함께 관리될 수 있도록 데이터베이스를 구축 • Dicom 헤더에 포함된 환자 정보를 제거

구분	내용	
		• 영상 내에서 Tagging 혹은 기기별 영상 레이아웃에 의해 포함되는 환자 식별 정보를 제거
	초음파 영상	자동화 툴에서 제거되지 못한 상하좌우에 포함된 식별정보 및 노이즈를 수동으로 제거
	CT	추가 없음
	병리 이미지	추가 없음

2.4 획득 및 정제 기준

〈표 Ⅲ-84〉 데이터 획득 및 정제 기준

유형	수집기준	배제기준
초음파	• 병변을 대표하는 B-mode 영상 • 양성/악성 기준 - 양성: 생검 및 수술로 양성으로 확진된 병변 - 악성: 생검 및 수술로 악성으로 확진된 병변	• 병변 내에 인공물(Indicator)이 포함되어 있는 경우
Neck CT	• 갑상선암으로 진단받은 환자 중 임파선 전이가 의심되는 환자 대상 - 경부 림프절 병변 확인 및 위치 - 2mm 이하의 절편 두께를 가지는 조영증강 영상	• 움직임에 대한 인공음영이 있는 경우 • 적절한 조영증강이 되지 않은 경우
병리 이미지	• 적절히 smear 된 병리 슬라이드 • 양성/악성 기준 - 양성: 생검 및 수술로 양성으로 확진된 병변 - 악성: 생검 및 수술로 악성으로 확진된 병변	• 적절히 smear 되지 않거나 세포의 양이 충분하지 않은 경우

3 어노테이션/라벨링

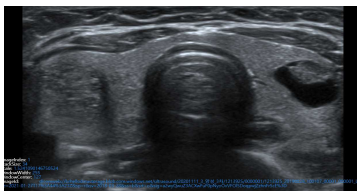
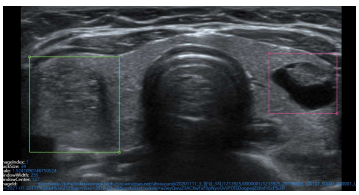
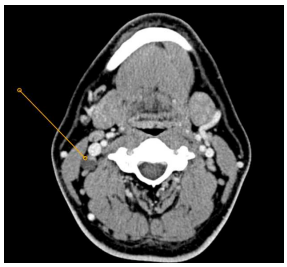
3.1 어노테이션 / 라벨링 절차

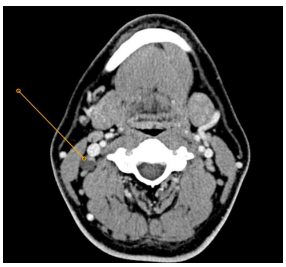
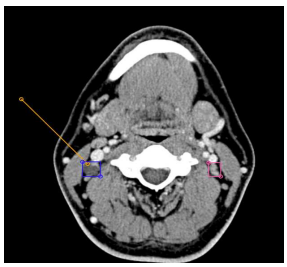
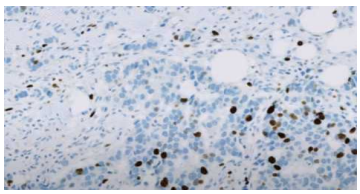
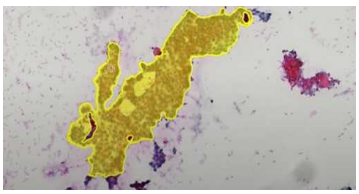
〈표 Ⅲ-85〉 어노테이션/라벨링 절차

구분	초음파	Neck CT	병리 이미지
1차 수행	• 수행인: 영상의학과 내분비 외과/내과, 이비인후과 의사 • 수행 내용: 병변부위 바운딩 박스 체크	• 수행인: 영상의학과 전문의 • 수행 내용: 병변부위 라인 체크	• 수행인: 병리학과 전문의 • 수행 내용: 이상부위 폴리곤 체크
2차 수행	해당사항 없음	• 수행인: 별도 교육을 받은 인력 • 수행 내용: 병변부위 바운딩 박스 체크	해당사항 없음

3.2 어노테이션 / 라벨링 기준

〈표 Ⅲ-86〉 어노테이션/라벨링 기준

<table>
<tr><th>유형</th><th colspan="2">어노테이션 항목</th><th>어노테이션 수행인</th></tr>
<tr><td rowspan="2">초음파</td><td colspan="2">• 병변부위 Bounding Box 체크</td><td rowspan="2">• 의사
– 내분비 내과
– 내분비 외과
– 영상의학과
– 이비인후과</td></tr>
<tr><td>[수행 전]
  </td><td>[수행 후]
  </td></tr>
<tr><td rowspan="2">Neck CT</td><td colspan="2">• 1차 작업
– 병변부위 라인 체크</td><td rowspan="2">• 의사
– 영상의학과 전문의</td></tr>
<tr><td>[수행 전]
  </td><td>[수행 후]
  </td></tr>
</table>

유형	어노테이션 항목		어노테이션 수행인
	• 2차 작업 – 병변부위 바운딩 박스 체크		• 데이터 레이블링 관련 교육을 수료한 의공학, 공학, 의학 등 전공자
	[수행 전] 	[수행 후] 	
병리 이미지	• 이상 부위 Polygon 형태 체크 • 갑상선 유두암 라벨링		• 의사 – 병리과 전문의
	[수행 전] 	[수행 후] 	

- 어노테이션 작업자 기준
  - 내부 전문가 그룹: 영상의학과, 내과, 외과, 이비인후과 의사
  - 외부 교육생: 데이터 레이블링 관련 교육을 수료한 의공학, 공학, 의학 등 전공자

3.3 어노테이션 / 라벨링 교육

〈표 Ⅲ-87〉 어노테이션/라벨링 작업자 교육

교육 구분	내용	수행 기간	비고
기본 교육	• 암종의 이해 • 데이터 형태에 대한 이해 • 어노테이션 도구 교육 • 실습	• 암종별 1일 • 반복 수행	온/오프라인
심화 교육	• 암종의 세부 교육 • 데이터 어노테이션 방법 • 데이터 어노테이션 결과 평가 방법 • 실습	• 암종별 1주 • 반복 수행	온 /오프라인

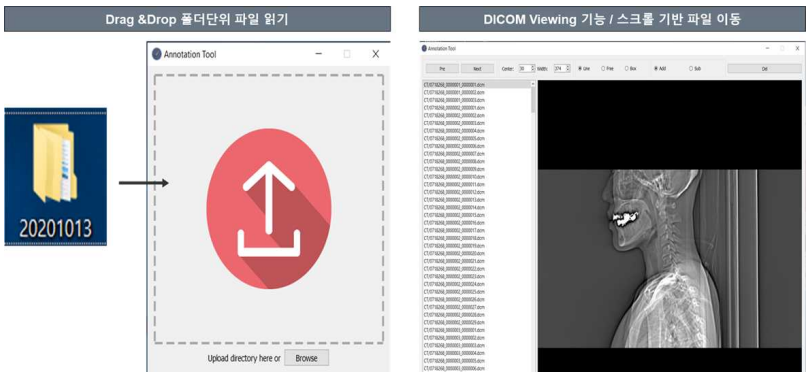
3.4 어노테이션 / 라벨링 도구 및 사용법

1) 저작도구

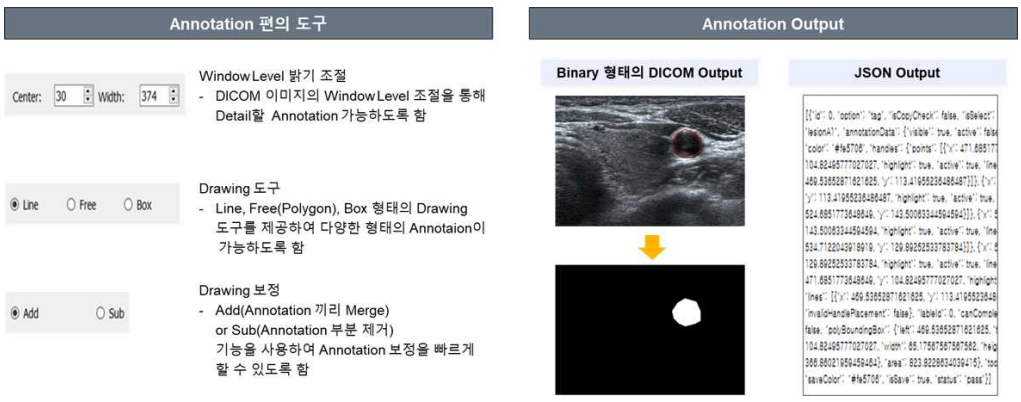
- 학습용 데이터 저작도구는 K-TIRADS 어노테이션, 종양영역 경계 어노테이션, 자동 어노테이션 결과 저장 등을 제공하는 경북대학교 산학협력단에서 개발한 Window/Python 기반 프로그램을 사용함

2) 기능

- 폴더 읽기 / DICOM VIEWING 기능
- Annoation Drawing 편의 및 Ouput

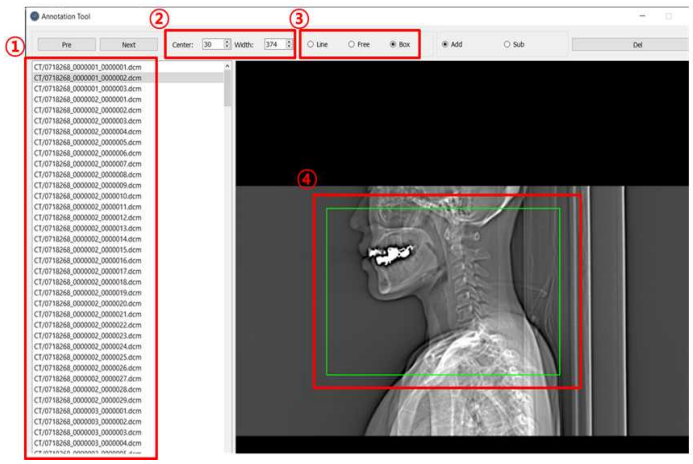


[그림 Ⅲ-100] 학습용 데이터 저작도구 폴더 읽기 / DICOM VIEWING 기능



[그림 Ⅲ-101] 학습용 데이터 저작도구 Annoation Drawing 편의 및 Ouput 기능

3) 절차



[그림 Ⅲ-102] 어노테이션 작업 화면 예시

4 데이터 검수

4.1 검수 절차

1) 초음파

- 1차: (어노테이션) 익명화 소프트웨어를 사용하여 US에서 개인 정보를 제외한 영상영역 추출 후 갑상선 초음파를 시행하는 내분비 내과, 내분비 외과, 영상의학과, 이비인후과 의사가 직접 영상에서 결절들을 사각형 bounding box 형태로 annotation 시행
- 2차: (검수) 영상의학과 전문의에 의해 100% 전수 교차 검증 (의견이 다를 경우 최종 검수자가 병리 결과 및 EMR, PACS 데이터를 확인하여 최종 결정함)

2) Neck CT

- 1차: (프리어노테이션)영상의학과 전문의가 병변 부위 라인 표시후 클라우드 서버에 업로드
- 2차: (어노테이션) 교육생이 bounding box 형태로 annotation 시행
- 3차: (검수) 2명의 영상의학과 전문의가 100% 전수 검사 교차 검증 (의견이 다를 경우 최종 검수자가 병리 결과 및 EMR, PACS 데이터를 확인하여 최종 결정함)

3) 병리 이미지

- 1차: (어노테이션) 병리과 전문의가 polygon 툴로 segmentation
- 2차: 병리과 전문의가 100% 전수 교차 검증 (의견이 다를 경우 최종 검수자가 병리 결과 및 EMR, PACS 데이터를 확인하여 최종 결정함)

4.2 검수 기준

1) 데이터 획득 단계(원본추출)

- PACS에서 영상추출을 위한 환자 대상자선정 및 추출 리스트 검토 2인 이상 확인 (암종별 정보, 환자정보 누락은 없는지, 정확한 검사 종류,날짜,시간 등)
- PACS에서 익명화하여 추출한 영상 파일을 원내 DATA서버로 옮겨 데이터 적절성 교차 확인 (DICOM head 정보, 영상의 누락은 없는지, 매핑정보(new ID)
- EMR ID 와 NEW ID의 매핑정보 재확인 후 클라우드 서버에 업로드

2) 2차 검수

- 2차 검수 시 전문의 의사가(초음파 및 CT – 영상의학과, 병리 이미지 – 병리과) 교차 검증 (의견이 다를 경우 최종 검수자가 병리 결과 및 EMR, PACS 데이터를 확인하여 최종 결정함)

5 데이터 활용 방안

5.1 학습 모델

- 학습 모델 개발 계획

〈표 Ⅲ-88〉 갑상선암 AI 학습용 데이터 인공지능 학습 모델 개발 계획

유형	모델 개발 계획	학습 분배량
초음파	• 인공지능 모델 개발 목표 – 초음파 이미지의 상의 결절 유무와 위치를 판단 – 결절의 감별 진단(양성/악성) • AI 베이스 라인 모델	* 학습:시험:검증 비율은 환자수를 기준으로 8:1:1로 구성되며, 이미지 장수는 차이가 발생할 수 있음.

유형	모델 개발 계획	학습 분배량
	- 객체 분류 및 탐지 분야에서 SOTA 결과를 보여주고 있는 Inception-v3 알고리즘 및 FASTER R-CNN 알고리즘을 우선적으로 적용하여 진행 - 수집 데이터에서 성능 향상을 위한 모델 개선, 신규 모델 적용을 수행하며, 목표 성능을 상회하는 모델에 대한 소스코드 등을 공개함	
Neck CT	• 인공지능 모델 개발 목표 - CT 기반 전이 유무 판단 및 전이 위치 추정 AI모델 개발 • AI 베이스 라인 모델 - 객체 분류 및 탐지에서 일반적으로 사용되는 VGG, Inception 등을 우선 적용하여 성능 확인 - 수집 데이터에서 성능 향상을 위한 모델 개선, 신규 모델 적용을 수행하며, 목표 성능을 상회하는 모델에 대한 소스코드 등을 공개함	* 악성 환자를 대상으로 조직 전이 여부를 라벨링 * 학습:시험:검증 비율은 환자수를 기준으로 8:1:1로 구성되며, 이미지 장수는 차이가 발생할 수 있음.
병리 이미지	• 인공지능 모델 개발 목표 - 세침흡입 세포 병리 이미지를 활용하여, 악성종양세포 세포의 유무 및 위치를 찾는 AI 모델 개발 - 세침흡입 세포 병리 이미지를 활용하여, 악성종양세포의 악성 정도를 파악하는 AI 모델 개발 • AI 베이스 라인 모델 - 객체 분류 및 탐지 분야에서 SOTA 결과를 보여주고 있는 VGG-19, Inception-v3 알고리즘을 우선적으로 적용하여 진행 - 수집 데이터에서 성능 향상을 위한 모델 개선, 신규 모델 적용을 수행하며, 목표 성능을 상회하는 모델에 대한 소스코드 등을 공개함	* 학습:시험:검증 비율은 환자수를 기준으로 8:1:1로 구성되며, 이미지 장수는 차이가 발생할 수 있음.

5.2 서비스 활용 시나리오

〈표 Ⅲ-89〉 갑상선암 AI 학습용 데이터 서비스 활용 시나리오

과제명	데이터 유형	응용서비스
갑상선암 AI 학습용 이미지 데이터 구축	초음파	• 결절 유무 판단 • 결절의 위치 표기 • 결절 감별 진단(양성/악성)
	Neck CT	• 전이 유무 판단 • 전이 위치 표기
	병리 이미지	• 악성종양세포 유무 판단 • 악성종양세포 위치 표기 • 악성 정도 판단

제10장 유방암 AI 학습데이터

1 데이터 정보 요약

1.1 가이드 분류

대분류	이미지	중분류	헬스케어	소분류	DICOM, MRI

1.2 데이터 정보

데이터 이름	유방암 AI 학습데이터
활용 분야	의료
데이터 요약	유방초음파, 유방촬영 및 유방MRI

1.3 데이터 구축 개요

1.4 구축 목적

- 종양 감별 진단 및 잔존암 유무 예측이 가능한 의료용 인공지능 학습 모델 개발을 위한 멀티모달 유방 영상 의료데이터베이스 구축
- 다양한 유방 영상의 멀티모달을 기반으로 AI 모델의 진단 정확도 향상

1.5 활용 분야

- 유방초음파에서 종양 양성, 악성 판별 모델 개발
- 유방영상의 다양한 멀티모달을 활용한 종합적 진단 모델 개발
- 유방 영상의 잔존암 유무 예측 모델 개발

1.6 유의 사항

- 환자의 의료정보가 포함되어 있는 의료 데이터(원시 데이터)는 승인된 연구자 이외에는 접근이 불가능함
- 보건복지부의 보건의료 데이터 활용 가이드라인에 따라, 데이터 활용 및 제 3자의 배포를 위해서는 해당 의료 기관의 데이터 심의기관의 허가를 받아야 함

2 데이터 획득 및 정제

2.1 원시 데이터 선정

- DICOM 이미지 655,838장

〈표 Ⅲ-91〉 원시데이터 유형 및 목표 건수

유형	구축 건수(건)		
	분류	환자 수	영상 수
유방 초음파	양성	600명	600장
	악성	600명	600장
유방 촬영술	잔존암이 없는 환자	150명	600장
	잔존암이 있는 환자	150명	600장
유방 MRI	잔존암이 없는 환자	150명	326,425장
	잔존암이 있는 환자	150명	326,413장

2.2 규제관련 사항

- 원시 데이터 획득을 위해 경북대병원의 IRB 승인 획득
- 환자 비식별화 및 익명화 작업을 수행하여 의료 관련 개인정보보호법에 침해되지 않도록 처리

2.3 획득 및 정제 절차

〈표 Ⅲ-92〉 데이터 획득 및 정제 절차

구분	내용	
데이터 획득	• 경북대병원의 의료데이터베이스(EMR, PACS)에서 유방암 환자의 유방초음파, 유방촬영술, 유방MRI 데이터를 획득	
데이터 정제	공통 사항	• 익명화 SW를 이용해 환자 ID, 영상 ID를 수집 데이터베이스에 의해 관리되는 기준 ID로 치환하고, 임상 데이터와 함께 가공 기관 및 관련 정보가 함께 관리될 수 있도록 데이터베이스를 구축 • Dicom 헤더에 포함된 환자 정보를 제거 • 영상 내에서 Tagging 혹은 기기별 영상 레이아웃에 의해 포함되는 환자 식별 정보를 제거
	유방초음파	• B-Mode 영상 외 도플러 영상 등 추가 진단용 영상은 제외함 • 영상 내 화살표나 길이 측정 등에 의한 태깅 정보가 포함된 B-Mode 영상은 제외함 • 영상 내 가장 자리에 기기에 따른 표기가 나타난 영역은 Cropping을 통해 제외함 • 비식별화 완료된 초음파 영상의 크기는 변환하지 않으며, AI 모델 구현에 따라 활용할 수 있도록 함
	유방촬영술	• 유방 촬영 방향에 대한 영상내 표기 정보는 제거하며, 유방 촬영 방향 정보는 DICOM 헤더에 저장하도록 함 • 유방촬영술 촬영 기준에 맞지 않는 영상은 제외함
	유방MRI	• 조영제 사용 전 영상과 사용 후 영상에 공통된 어노테이션을 사용할 수 있도록 모든 영상의 공간정보 일관성 확인

2.4 획득 및 정제 기준

〈표 Ⅲ-93〉 데이터 획득 및 정제 기준

유형	수집기준	배제기준
유방 초음파	• 병변을 대표하는 B-mode 영상 • 양성/악성 기준 – 양성: 생검 및 수술로 양성으로 확진된 병변 – 악성: 생검 및 수술로 악성으로 확진된 병변	• 병변 내에 인공물(Indicator)이 포함되어 있는 영상 또는 움직임에 대한 인공음영이 있는 경우
유방 촬영술	• 한 명당 4장의 표준 영상수를 충족하며 한국의료영상품질관리원의 기준을 충족하는 영상	• 움직임에 대한 인공음영이 있는 경우
유방 MRI	• 필수 Sequence (T2, 조영 전 T1, 조영 후 T1, subtraction영상)이 있고, 한국의료영상품질관리원의 기준을 충족하는 영상	• 인공음영이 병변의 segmentation을 방해하는 경우

3 어노테이션/라벨링

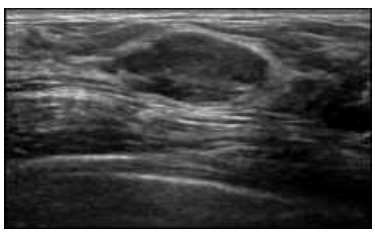
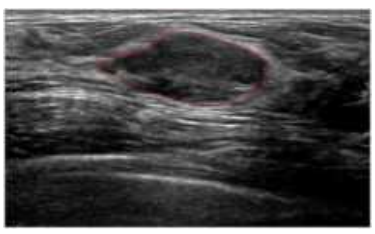
3.1 어노테이션 / 라벨링 절차

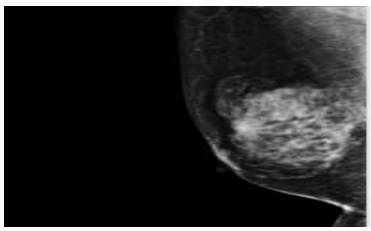
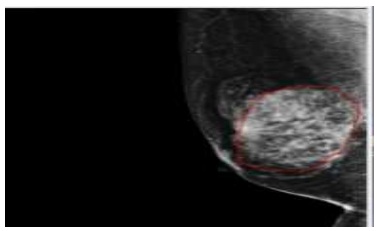
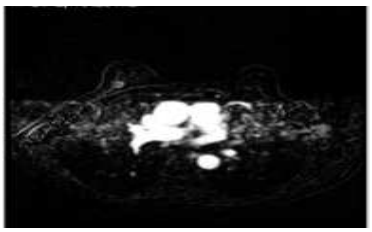
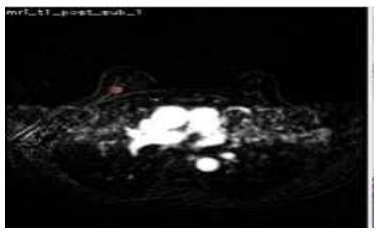
〈표 Ⅲ-94〉 어노테이션/라벨링 절차

구분	유방초음파	유방촬영술	유방MRI
1차 수행	• 수행인: 영상의학과 전문의 • 수행 내용: 양성/악성 분류	• 수행인: 영상의학과 전문의 • 수행 내용: 잔존암 유무 분류	• 수행인: 영상의학과 전문의 • 수행 내용: 잔존암 유무 분류
2차 수행	• 수행인: 별도 교육을 받은 데이터엔지니어 • 수행 내용: 병변부위 Polygon 세그멘테이션	• 수행인: 별도 교육을 받은 데이터엔지니어 • 수행 내용: 병변부위 Polygon 세그멘테이션	• 수행인: 별도 교육을 받은 데이터엔지니어 • 수행 내용: 병변부위 Polygon 세그멘테이션

3.2 어노테이션 / 라벨링 기준

〈표 Ⅲ-95〉 어노테이션/라벨링 기준

<table>
<tr><th>유형</th><th colspan="3">어노테이션 항목 및 기준</th><th>어노테이션 수행인</th></tr>
<tr><td rowspan="4">유방 초음파</td><td colspan="3">• 1차 작업
- 양성/악성 분류</td><td rowspan="3">• 영상의학과 전문의</td></tr>
<tr><td rowspan="2">기준</td><td>양성</td><td>• 절제 생검으로 양성으로 진단된 경우
• 조직검사로 양성으로 진단된 후 1년 이상 추적검사에서 변화가 없는 경우
• 양성추정 병변으로 2년이상 추적검사에서 변화가 없는 경우</td></tr>
<tr><td>악성</td><td>• 조직검사 및 수술로 악성(암)으로 진단된 경우</td></tr>
<tr><td colspan="3">• 2차 작업
- 병변부위을 포함하여 최소한의 margin을 두고 polygon 세그멘테이션을 실시</td><td rowspan="2">• 별도 교육을 받은 데이터엔지니어</td></tr>
<tr><td colspan="3">   </td></tr>
</table>

유형	어노테이션 항목 및 기준	어노테이션 수행인
유방 촬영술	• 1차 작업 – 잔존암 유무 분류	• 영상의학과 전문의
	• 2차 작업 – 병변부위을 포함하여 최소한의 margin을 두고 polygon 세그멘테이션을 실시	• 별도 교육을 받은 데이터엔지니어
	 	
유방 MRI	• 1차 작업 – 잔존암 유무 분류	• 영상의학과 전문의
	• 2차 작업 – 병변부위을 포함하여 최소한의 margin을 두고 polygon 세그멘테이션을 실시	• 별도 교육을 받은 데이터엔지니어
	 	

- 어노테이션 작업자 기준
  - 내부 전문가 그룹: 영상의학과 전문의
  - 외부 교육생: 별도 교육을 받은 데이터엔지니어

3.3 어노테이션 / 라벨링 교육

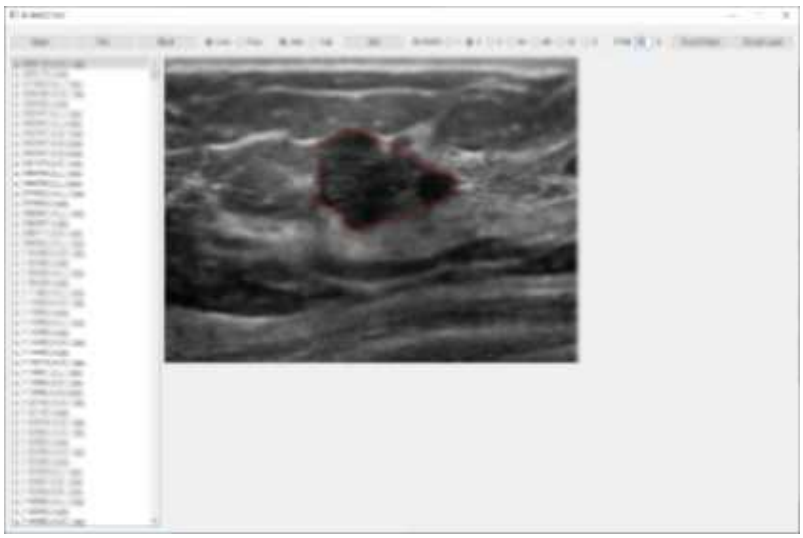
- 어노테이션을 수행하는 전문의 및 외부 수행자를 대상으로 어노테이션 툴(저작 도구) 사용 방법, 어노테이션 방법 등 교육 진행

〈표 Ⅲ-96〉 어노테이션/라벨링 작업자 교육

교육 구분	내용	수행 기간	비고
기본 교육	• 암종의 이해 • 데이터 형태에 대한 이해 • 어노테이션 도구 교육 • 실습	• 암종별 1일 • 반복 수행	온/오프라인
심화 교육	• 암종의 세부 교육 • 데이터 어노테이션 방법 • 데이터 어노테이션 결과 평가 방법 • 실습	• 암종별 1주 • 반복 수행	온 /오프라인

3.4 어노테이션 / 라벨링 도구 및 사용법

- 내부 전문가용/공개용 저작도구 개발 방안
  - 유방암 관련 내부 전문가용(영상전문의, 준전문가)으로 의료 특수성을 감안한 저작도구 개발
  - 유방초음파, 유방촬영술 및 유방MRI의 어노테이션이 가능하고 영상 정보를 유지하면서 DICOM 포맷으로 어노테이션 정보가 저장되도록 구현
  - 어노테이션에 대한 검토 및 수정이 가능하도록 데이터 저장/로드, 어노테이션 보정 기능 구현
  - 진단 AI서비스와 연계된 다양한 영상 모달리티 저작도구 개발 및 공개
  - 경북대학교 산학협력단에서 설치형 프로그램으로 개발



[그림 Ⅲ-103] 학습용 데이터 저작도구

4 데이터 검수

4.1 검수 절차

〈표 Ⅲ-97〉 데이터 검수 절차

구분	수행주체	내용
1차 검수	경북대학교	• 태깅 확인 • 데이터 구성 및 폴더구조 확인
2차 검수	유방영상의학과 전문의	• 모든 부분이 정확히 어노테이션 된 경우 approve • 어노테이션이 정확하지 않을 경우 직접 drawing하여 approve, 부정확한 부분에 대한 피드백 제공 • 전문의 2인 이상 교차 검증

4.2 검수 기준

〈표 Ⅲ-98〉 데이터 검수 기준(1차 검수)

구분	검수기준
공통	• 어노테이션 영역이 병변 영역을 90%이상 포함한다고 판단이 될 때 승인 • 검수를 통과하지 못하는 경우에는 기존의 어노테이션을 삭제하고 검수자가 다시 drawing한다.
자주 발생하는 오류 사례	• 병변이 암인 경우, 주변으로 딸결절 (daughter nodule)이나 제자리암 (in situ cancer)을 포함하지 않은 경우: 검수자가 직접 상기 병변을 포함하여 polygon 세그멘테이션을 수정 혹은 수행한다. • 유방초음파에서 병변 후방 저에코를 병변의 영역을 넘어서 포함할 때: 검수자가 직접 병변 후방 저에코 영역을 제외하고 병변 영역만 polygon 세그멘테이션을 수정 혹은 시행한다.
검수자간 의견이 다를 경우 처리절차	• 검수자간 의견이 다를 경우에는 내부 검수자인 최종 검수자가 병리 결과 및 EMR, PACS 정보를 확인하여 최종 어노테이션 영역을 결정함.

- 전문가 교차 검증 방안(2차 검수)
  - 수집 의료 데이터의 수집 환경 (장비, 영상 획득 프로토콜 등), 판독 일치성과 레이블링 데이터의 정확성 및 유효성의 정도를 고려한 교차 검증 프로세스 기반으로 수행

〈표 Ⅲ-99〉 전문가 교차 검증 방안(2차 검수)

구분	검수전문가 수	내용
유방촬영술	3인	외부 전문가를 활용하여 1차, 2차 검수을 순차적으로 거친 어노테이션영역을 확인하여 내부 검수자가 병리 결과 및 EMR, PACS 정보를 확인하여 최종 어노테이션 영역을 결정함.
유방초음파	2인	내부 전문가의 독립된 검수 영역 중 중첩된 영역을 최종 어노테이션 영역으로 결정함.
유방MRI	2인	외부 전문가를 활용하여 1차 검수를 거친 어노테이션 영역을 확인하여 내부 검수자가 병리 결과 및 EMR, PACS 정보를 확인하여 최종 어노테이션 영역을 결정함.

5 데이터 활용 방안

5.1 학습 모델

〈표 Ⅲ-100〉 유방암 AI 학습용 데이터 인공지능 학습 모델

유형	모델 개발 계획	학습 분배량
유방 초음파	• 인공지능 모델 개발 목표 - 초음파 기반 유방종양 양성/악성 감별진단 AI모델 개발 - 초음파 영상을 입력으로 하여, 영상 특징 학습 후 취합하여 양성, 악성을 구분하는 Binary Classification 모델을 개발 • AI 베이스 라인 모델 - 초음파 감별진단 AI 베이스라인 모델으로 Classification에서 일반적으로 사용되는 VGG-16, Residual Network, GooLeNet, DenseNet을 우선 적용함 - 수집 데이터에서 성능 향상을 위한 모델 개선, 신규 모델 적용을 수행하며, 목표 성능을 상회하는 모델에 대한 소스코드 등을 공개함	1. 학습셋 - 총 1,000장 - 양성: 500장 - 악성: 500장 2. 검증셋 - 200장 - 양성: 100장 - 악성: 100장

유형	모델 개발 계획	학습 분배량
유방 촬영술	• 인공지능 모델 개발 목표 - 유방촬영술 및 유방MRI 기반 잔존암 유무 예측 모델 개발 - 유방 촬영술 및 유방MRI 영상의 ROI 영상을 입력으로 하여, 영상 특징 학습 후 취합하여 잔존암 유무 판별 (pCR classification)을 하는 Binary Classification 모델을 개발함 • AI 베이스 라인 모델 - AI 베이스라인 모델로 classification에서 일반적으로 사용되는 VGG-16, Residual Network, GooLeNet, DenseNet을 우선 적용함 - 수집 데이터에서 성능 향상을 위한 모델 개선, 신규 모델 적용을 수행하며, 목표 성능을 상회하는 모델에 대한 소스코드 등을 공개함 - 단일 모달리티 기반 AI 베이스라인 모델 외에도 멀티 모달리티의 영상특징을 취합해 잔존암 유무를 판별하는 AI 모델을 개발, 평가하며 각각 모델의 결과를 제시함.	1. 잔존암이 없는 환자 - 학습셋: 100명 - 검증셋: 50명 2. 잔존암이 있는 환자 - 학습셋: 100명 - 검증셋: 50명
유방 MRI		1. 잔존암이 없는 환자 - 학습셋: 100명 - 검증셋: 50명 2. 잔존암이 있는 환자 - 학습셋: 100명 - 검증셋: 50명

5.2 서비스 활용 시나리오

〈표 Ⅲ-101〉 유방암 AI 학습용 데이터 인공지능 학습 모델

과제명	데이터 유형	응용서비스
유방암 AI 학습용 이미지 데이터 구축	유방초음파	• 유방종양 감별 진단(양성/악성)
	유방촬영술	• 잔존암 여부 예측 모델
	유방MRI	• 병변 검출 및 탐지 • 폐 결절과 임파선의 양성/악성 감별 • 악성 폐 결절과 악성 임파선 검출

제11장 주행환경 정적객체 데이터

1 데이터 정보 요약

1.1 가이드 분류

대분류	이미지	중분류	비전	소분류	JPG

1.2 데이터 정보

데이터 이름	주행환경 정적객체 데이터
활용 분야	자율주행 영상 인식 기술 개발, 자율주행을 위한 정밀 지도 제작 자동화 기술 개발 등
데이터 요약	수도권 및 광역시, 고속도로, 국도 등 전국 다양한 지역의 도로 주행환경에서 정적객체(차선/횡단보도 및 신호등/교통표지판)에 대한 영상 및 레이블링 데이터를 제공. 차선 및 횡단보도의 경우 폴리곤/폴리라인으로 라벨링이 되어 있으며, 신호등 및 표지판 데이의 경우 bounding box 형태로 라벨링이 되어 있음

1.3 데이터 구축 개요

〈표 Ⅲ-102〉 단계별 데이터 구축 개요

단계	세부절차	설명
수집	영상 단독 데이터 수집	차량 주행 영상 획득 영상 클라우드 업로드
	다중 센서 데이터 수집	데이터 수집 장비 구성 데이터 수집 조건 분석 (위치, 환경, 기상) 데이터 수집 동선 및 일정 계획 수집 장비 캘리브레이션 주행 간 데이터 수집 데이터 유효성 검증
정제	영상 단독 데이터 정제	데이터 취득 조건에 따른 영상 추출 및 파일명 생성 이미지 추출(JPG)
	다중 센서 데이터 정제	데이터의 객체 포함 및 활용성 기반하여 큐레이션 후 시퀀스 화

단계	세부절차	설명
가공	가공 인력 양성 교육	신규 작업자 레이블링 교육 훈련
	데이터 가공	수집 및 정제된 영상을 크라우드 소싱을 통해 프레임별로 객체 정보를 레이블링 in-house 인력을 활용한 레이블링
검수	탐지 정확도 검수	라벨링 데이터 검출 적절성 여부(오검출, 미검출, 과검출 여부 판단)
	위치 정확도 검수	라벨링 경계선(정확도) 적절성 여부

1.4 구축 목적

- 자율주행 기술의 고도화를 위해서는 영상인식 AI 기술의 발전이 필수적이며 이를 위해 다양한 영상 데이터 구축이 필요
- 자율주행 정적객체 (차선/횡단보도 및 신호등/교통표지판) 인식을 위한 대규모 데이터 셋을 구축하여 공개하고, 해당 데이터를 활용한 프로토타입 모델/서비스/코드를 공개

1.5 활용 분야

- 차선 인식을 통한 자율주행 차량의 횡방향 측위 기술 고도화
- 신호등 및 표지판 인식을 통한 도심 자율주행의 인지 알고리즘 고도화
- 차선/횡단보도/신호등/표지판 인식을 통한 지도 제작 자동화 기술에 활용

1.6 유의 사항

1) 데이터 3법 시행령에 따른 블랙박스 영상 사용 합법성 검토
   - 블랙박스 영상은 고객의 개인정보에 해당하므로 기존의 정보 수집 목적인 '사고원인 규명 및 신속처리'에 관한 것이 아니라면 활용 불가능하나, 가명 처리할 경우 통계조사, 과학적 연구, 공익적 기록보관에 한 해 활용 가능
   - 과학적 연구의 경우 상업 목적의 연구도 포함
2) 개인정보 비식별화 처리 진행
   - 개인정보 및 초상권에 문제가 될 수 얼굴, 차량번호판 영역에 대해 블러 (blur) 처리 진행

[그림 Ⅲ-104] 얼굴 비식별화 예시

[그림 Ⅲ-105] 차량 번호판 비식별화 예시

2 데이터 획득 및 정제

2.1 원시 데이터 선정

- 원시데이터의 수집환경의 경우, 1.6절에서 설명되었듯이 짧은 사업 수행기간 내에 다양하고 많은 시간 데이터를 비용 효율적으로 수집하기 위해, 각 수집기관 별로 보유 인프라에 맞게 수집 환경조건을 다르게 해서 진행
- 3개의 수집 환경 (쏘카, 모빌테크, 라이드플럭스)으로 구분되며 각 수집 기관별 상세 내용 방식은 다음과 같음

① 쏘카 데이터 (전국)

- (데이터 비편향성 확보) 데이터 수집 일시, 위치에 따라 데이터를 수집하였으며 차선/횡단보도의 정적객체는 1.5절 데이터 통계에 따라 비편향적으로 데이터 셋을 구성
- 데이터 수집 위치는 서울/수도권, 고속화도로, 지방도로로 구분하여 데이터를 수집
- 데이터 수집 시간: 주간/야간대로 시간을 나누어 수집
- 데이터 수집 기상: 우천 데이터 확보를 위해 장마철인 7~8월 데이터 및 기상청에서 악천후가 기록된 날짜의 데이터 수집
- 지역별 수집 요약

〈표 Ⅲ-103〉 쏘카 데이터 지역별 수집 데이터 정보 요약

지역	합계 : 이미지 수 (1Hz 기준)	획득영상길이 (시간 환산)
Busan	51550	14
Chungcheongbuk-do	35959	10
Chungcheongnam-do	71570	20
Daegu	8557	2
Daejeon	33146	9
Gangwon-do	78776	22
Gwangju	11915	3
Gyeonggi-do	1875385	521
Gyeongsangbuk-do	26931	7
Gyeongsangnam-do	61549	17
Incheon	183344	51
Jeju-do	19808	6
Jeollabuk-do	16177	4
Jeollanam-do	12864	4
P'yongan-namdo	62	0
Seoul	395955	110
Ulsan	10682	3
총합계	2894230.06	804

② 모빌테크 데이터 (수도권 지역)

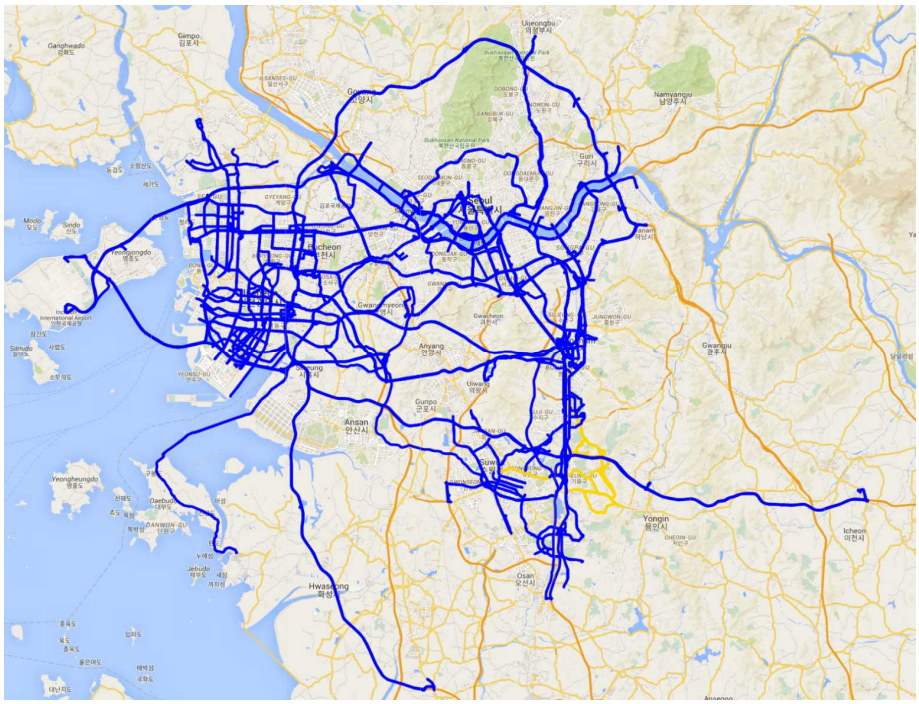
- 원시 데이터가 AI 학습에 유용한 정보를 담고있는지 판단하기 위하여 (1) 지역 조건, (2) 기상 및 환경 조건, (3) 객체 포함 여부를 종합적으로 판단

  • 원시 데이터 선정 세부사항

〈표 Ⅲ-104〉 모빌테크 원시데이터 선정 세부사항

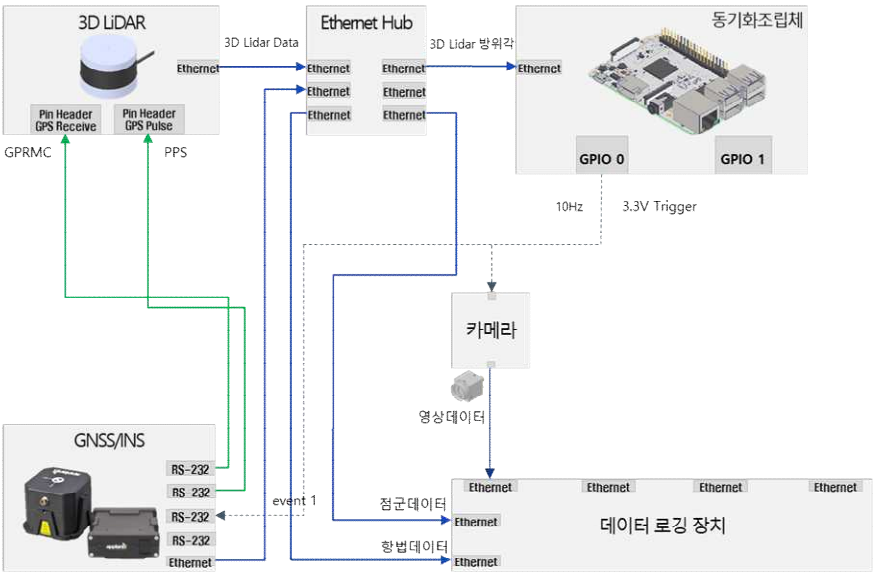
지역 조건	지역 조건을 평가하기 위하여 GNSS 좌표를 기준으로 중복되지 않는 지역을 원시데이터 선정 과정에 포함. 주요 도로 (고속도로 및 간선도로) 등으로 구분되는 도로 형태를 나누어 중복되지 않도록 함. 주요 도로를 우선적으로 수집, 그 뒤 8차선 이상 도로를 지역구역별로 수집하고, 최종적으로 지역 구별 4차선 이내 도로를 수집
시간 및 기상	기상 및 환경 조건을 평가하기 위하여, 맑은 날, 흐린 날, 우천, 안개, 일출, 일몰 등으로 나누어 구분하고 각 조건 별로 다양하게 포함 할 수 있도록 원시데이터를 선정
객체 포함 여부	수집 된 데이터 내에 정적 객체가 충분히 포함되어 있는지 여부를 평가함. 특히 시퀀스를 생성하기 위하여 정적 객체가 시퀀스에 포함되는지 여부를 평가하여 생성

- 수집 경로 발췌

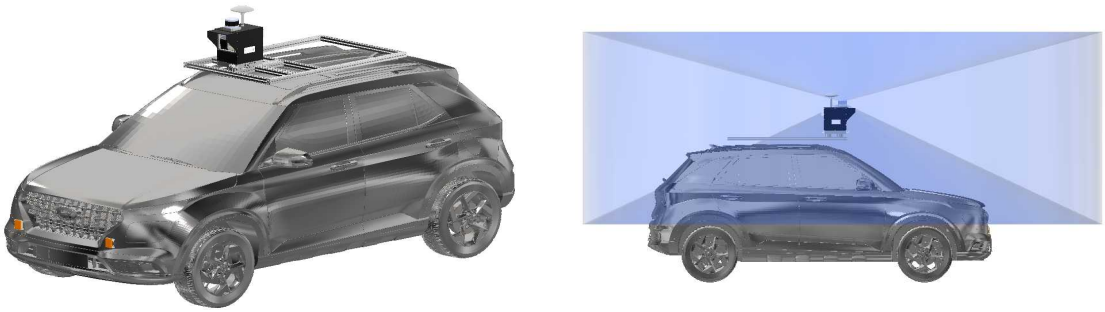


[그림 Ⅲ-106] 모빌테크 데이터 수집 경로

• 데이터 수집장치 정보



[그림 Ⅲ-107] 모빌테크 데이터 수집 장치 구성도



〈센서의 설치 위치〉 〈라이다 데이터의 시계〉

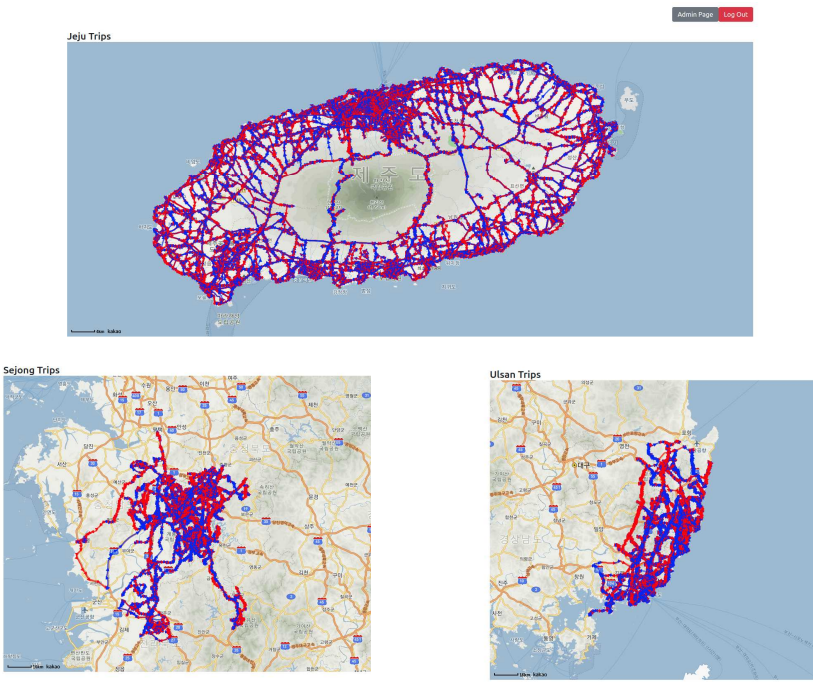
[그림 Ⅲ-108] 모빌테크 수집 차량 예시

[그림 Ⅲ-109] 모빌테크 수집 차량 실사

③ 라이드플럭스 데이터 (광역시, 고속도로, 국도 등 비수도권 지역)

- 원시 데이터 특성: 차량주행 중 전방을 향해 설치된 카메라 및 제반 센서로부터 획득된 정보
- 각 센서간의 시간적 동기화 및 센서 위치의 공간적 calibration 실시
- 주행환경 정적객체의 경우 취득하고 하는 객체 중 차선을 제외한 모든 객체(신호등, 횡단보도, 표지판)는 10분 이상 주행을 해도 한번도 발생하지 않을 수 있으므로, 수집목표량 대비 많은 수집시간과 정제과정이 필요
- 원시 데이터 선정 세부사항

〈표 Ⅲ-105〉 라이드플럭스 데이터의 원시 데이터 선정 세부사항

수집 지역	• 최대한 동일 도로를 반복하여 주행하지 않도록 사전에 수집 주행 경로 설정 (GNSS 좌표 및 도로 지도를 활용) • 자율주행에 관심도가 높은 지방 시/도(제주, 세종, 울산) 및 인근 지역 (대전, 청주, 공주, 양산, 부산, 경주 등)의 전국적 도로 환경 데이터 취득 
시간 및 기상	• 매일 오전 8시부터 오후 9시까지 데이터를 수집하여 주/야간, 일몰 및 일출에 대한 역광 데이터를 고르게 확보할 수 있도록 함 • 각 영상클립(메타데이터)별로 눈/비/안개 등의 악천후 유무를 체크
도로 환경	• 도로의 차선 정보 (편도 기준 1차선 이하, 2차선, 3차선, 4차선 이상)에 따라 데이터를 구분하여 확보 • 도로환경 (고속화도로, 도심도로, 교외도로)를 구분하여 확보

• 데이터 수집장치 정보

〈수집 차량 실사〉

〈트렁크 내부 (컴퓨팅/전원 장비 및 IMU)〉

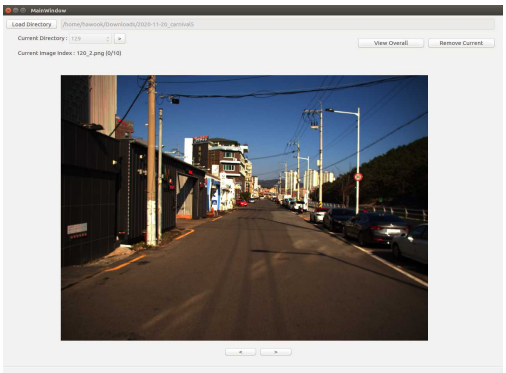
〈GNSS 안테나 설치〉

〈라이다/카메라 설치〉

[그림 Ⅲ-110] 라이드플럭스 데이터 수집장치 정보

• 수집/정제 도구

〈표 Ⅲ-106〉 라이드플럭스의 수집/정제 도구

수집 SW 구성	정제 SW 구성
	
• 보조석 인원이 주변상황 및 수집 중인 영상을 확인하면서 발생하는 이벤트 (횡단보도, 신호등, 표지판 출현 또는 희귀 데이터 출현) 기록 • 데이터 수집 환경 기록 • 40초 가량의 영상클립에서 적어도 정적 객체 하나 이상을 포함하도록 수집	• 수집된 영상을 1~2초 간격으로 재차 확인하면서 신호등/표지판 등의 객체 존재여부를 판단 • 객체가 없거나, 동일 영상 반복, 상태가 안좋은 영상 등을 삭제

2.2 규제관련 사항

- 관련 내용 없음

※ 본 데이터 가이드라인 상에는 규제 관련 내용이 없지만, 데이터 구축 시 규제 관련 내용 확인 및 검토 필요

2.3 획득 및 정제 절차

① 쏘카 데이터(전국)

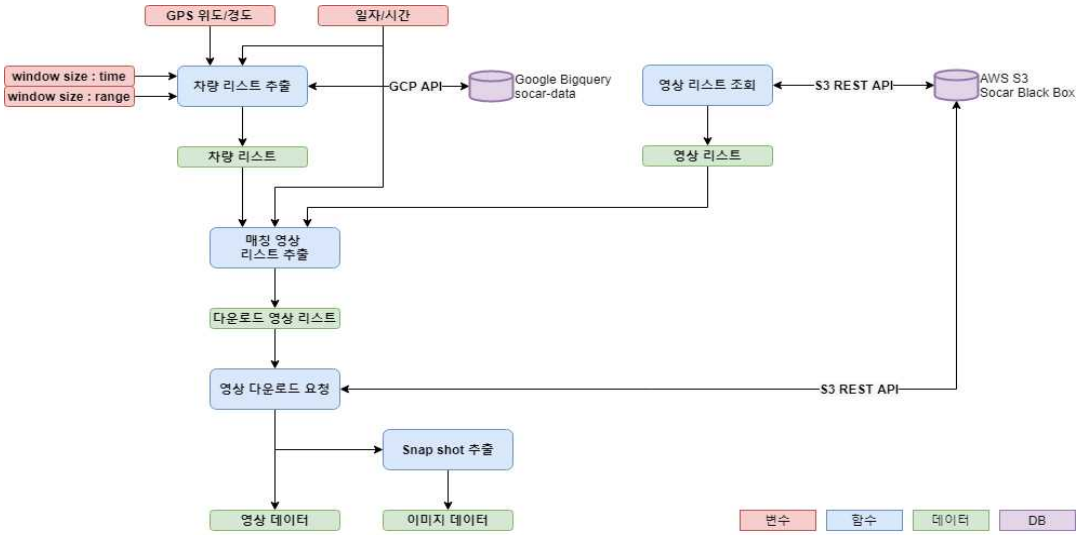
(1) (획득) 수집하고자 하는 지역의 위도, 경도 정보를 기준으로 지오펜스*을 설정하며 설정된 지오펜스에서 생성된 영상을 클라우드로부터 수집. 또한, 영상 생성 시간을 기준으로 클라우드로부터 영상을 수집

* 지오펜스 : GPS 정보를 이용한 가상의 울타리

(2) (정제) 수집한 영상을 AI학습데이터로 제공하기 위해 영상의 정보를 파일명으로 전달. 파일명에는 영상생성 일시, 주행 영상의 종류 구분, 전후방 카메라 구분으로 정보를 전달

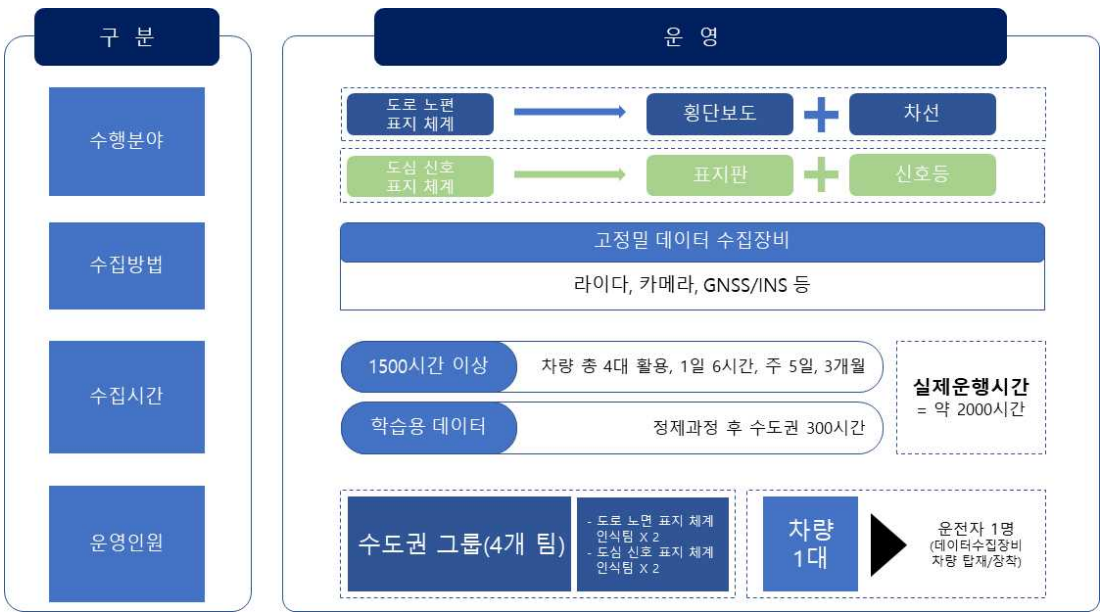
(3) 수집한 영상으로부터 JPG를 추출

(4) 주행영상 데이터 취득 다이어그램 (하단 그림 참조)

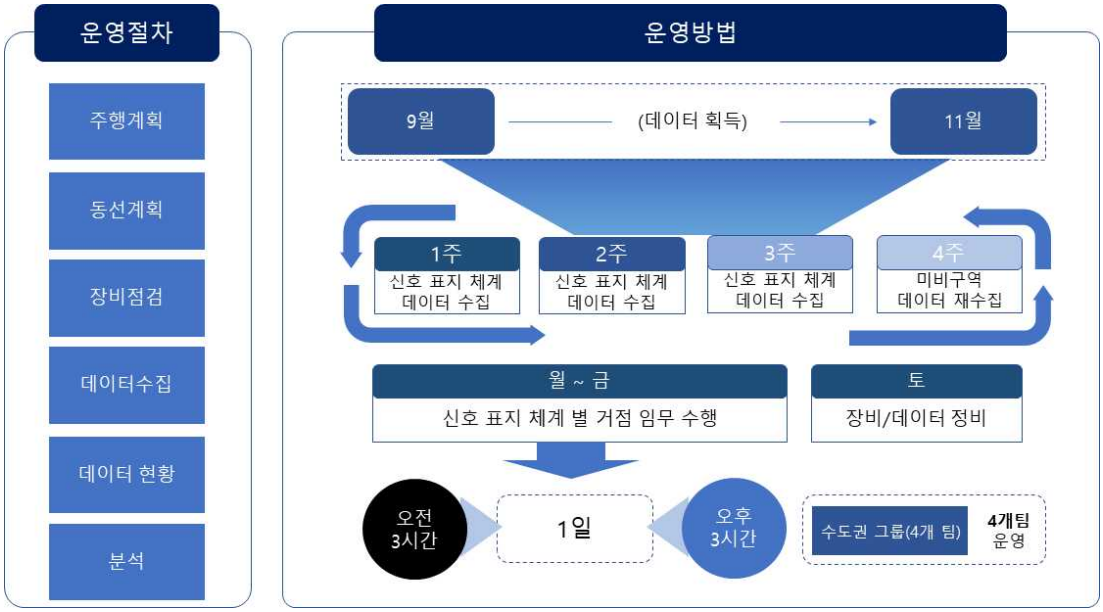


[그림 Ⅲ-111] 주행영상 데이터 취득 다이어그램(쏘카)

② 모빌테크 데이터 (수도권 지역)



[그림 Ⅲ-112] 데이터 수집 개요 (모빌테크)



[그림 Ⅲ-113] 데이터 수집 계획 (모빌테크)

③ 라이드플럭스 데이터 (광역시, 고속도로, 국도 등 비수도권 지역)

〈표 Ⅲ-107〉 데이터 수집 및 정제 절차 (라이드플럭스)

수집 절차	• 데이터 취득시 발생할 수 있는 위험성을 줄이고, 데이터 취득 효율화를 위하여 각 차량에 2명이 탑승하여 데이터 수집을 진행 • 조수석에 앉아있는 인원이 드물게 발생하는 데이터가 수집되는 상황을 체크할 수 있도록 함 • 조수석에 앉아있는 인원이 특이 데이터가 수집되는 상황마다 해당 시간대를 기록하면 자동으로 로깅된 데이터 중에서 원하는 타입의 데이터를 획득 가능 • 같은 구간을 지나더라도 날씨와 같은 주변 환경이 변하면서 발생하는 극한 상황(역광, 안개, 눈, 비, 야간)에 대해서도 별도로 체크를 해둠으로써 극한 상황에서의 인지 성능 테스트를 위한 레이블링 데이터 분류 • 차량 출고 전 가이드: 음주측정 → 메신저 단체방에서 실험 공유사항 확인 → 데이터 수집용 외장 SSD 확인 → 센서 렌즈 세척 → 차량 시동 및 PC 전원 공급 장치 가동 → 데이터 수집 프로그램 실행 • 차량 운행 종료 후 가이드: 운행 종료 시간 공유 → GNSS/INS 외부 기록계 중지 → 법인 핸드폰 주행경로 기록 저장 및 공유 → 데이터 수집 프로그램 및 차량 시동 OFF
정제 절차	• 수집 시 보조석 인원이 체크한 시점의 데이터를 자동으로 정제 • 정지가 오래 지속된 데이터 (e.g., 동일영상 반복), 영상 상태가 좋지 않은 데이터에 대한 정제 작업 • 시간적 편향성 방지를 위해 수집되는 데이터에서 1초 이상 간격으로 레이블링이 될 수 있도록 함 • LiDAR 센서의 rolling shutter 왜곡 현상 보정 작업 수행

2.4 획득 및 정제 기준

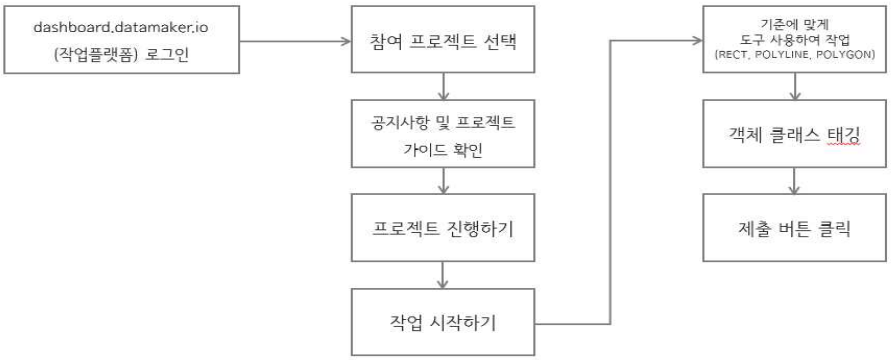
- 실 주행 영상에서 최대한 다양한 장소/시간/도로환경에서 영상 수집
- 차량이 오래 멈춰 있는 비디오 클립 제외 (e.g., 차량 움직임이 없을 경우 5~10Hz가 아닌 0.2Hz 간격으로 데이터 수집, 1분간 움직임이 전혀 없는 영상 클립 제거 등 수집기관별 상이)
- 동일 구간을 지속적으로 반복하며 촬영 금지 (e.g., 매일 주행 경로를 기록하고, 해당 경로를 최대한 회피하여 수집 계획 수립 등 수집기관별 상이)
- 희귀 원천 데이터 정제 방안
  - 신호등, 표지판, 횡단보도 등은 주행 중에 대부분 존재하지 않은 객체로, 일반 주행환경에서 1초의 동일 간격으로 영상을 캡쳐할 경우 평균 20%내외의 영상에만 객체가 존재 (10초 영상 시퀀스에서 1개의 표지판을 지날 경우에도 1~3개의 영상만 객체 존재)
  - 이는 RFP상 세부과제별 300시간의 원시데이터 영상을 1초 간격으로 균등 분할 할 경우, 추출가능한 1,080,000개의 원천데이터의 대부분이 객체가 포함되지 않는 문제가 생길 수 있음
  - 따라서 RFP상의 목표대비 130% 이상의 원시데이터를 초과 수집하며, 0.2~0.5초 등 짧은 간격의 영상 일지라도 신호등/표지판처럼 희소하게 발생하는 정적 객체의 외형 변화가 두드러지는 일부 데이터에 대해서는 원천데이터로 정제 허용 (하단 예시 그림 참조)
  - 또한 다양한 환경 (동적 객체 과제 공유 영상 등)에서의 탐지 정밀도 (precision)을 높이기 위한 데이터 확보를 위해 소량의 객체 없는 영상도 정상 원천 데이터에 포함

3 어노테이션/라벨링

3.1 어노테이션 / 라벨링 절차

1) 차선/횡단보도

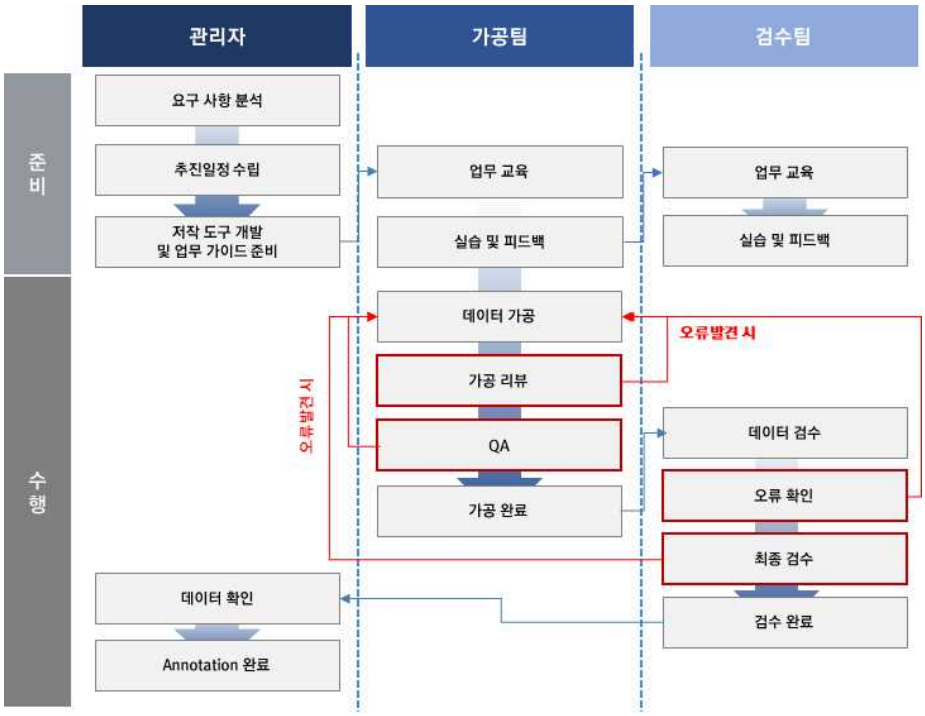
작업자 프로세스



[그림 Ⅲ-114] 어노테이션/라벨링 작업자 작업 프로세스

- 프로젝트 페이지에서 작업 시작 버튼 클릭시 할당된 영상(이미지) 프레임이 유저에게 노출됨
- 라벨링 해야 할 항목: 횡단보도(Polygon), 정지선(Polyline), 차선(Polyline)
- Polygon은 객체의 경계선을 따라 점을 찍어 그리고(Mouse Click, Touchpad Touch), Polyline은 차선의 중심부를 따라 Point를 찍어 그리는 차이가 있음
- Polygon, Polyline을 다 그린 후에는, 적절한 속성을 태깅함.(1차 속성: 차선, 정지선, 횡단보도)
- 각 클래스 부여가 완료되면, 각각의 클래스에 설정되어 있는 2차 속성들을 태깅함
- 차선의 경우 Color(흰색, 노란색, 파란색), Type(실선, 점선) 입력을 진행
- 모든 객체 라벨링이 완료되면 제출하기 버튼 클릭하며, 해당 작업물은 '작업완료' 형태로 변경되고, 다음 작업할 영상(이미지)을 호출함
- 이미지의 화질에 문제가 있거나, 어떠한 사유로 라벨링 할 수 없는 이미지를 할당 받은 경우, 제출하기 대신 보류 버튼을 눌러서 처리하며, '보류' 형태로 변경됨. 해당 작업물은 검수자의 판단하에 사용할 수 없는 이미지의 경우 '폐기' 처리

2) 신호등/표지판



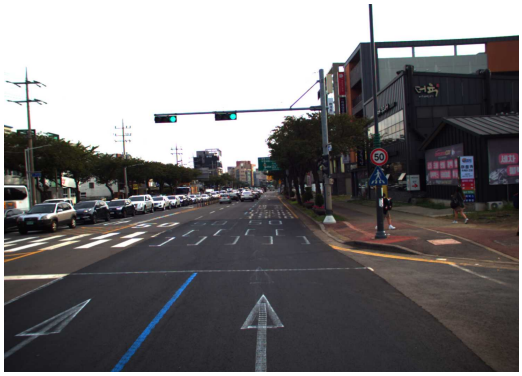
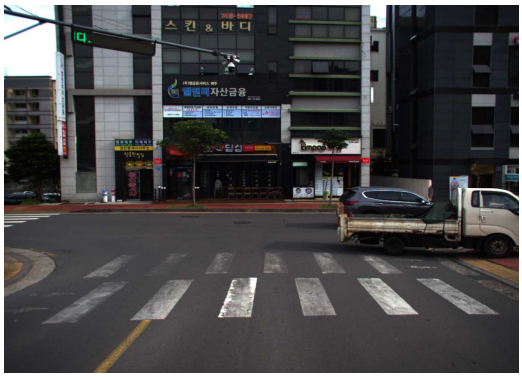
[그림 Ⅲ-115] 데이터 어노테이션 절차

3.2 어노테이션 / 라벨링 기준

1) 차선/횡단보도

〈표 Ⅲ-108〉 차선/정지선/횡단보도 어노테이션/라벨링 기준

주행 환경		묘사	사진
번호	레이블명		
①	차선	- 흰색 실선, 점선 - 노란색 실선, 점선 - 파란색 실선, 점선	

주행 환경		묘사	사진
번호	레이블명		
②	정지선	- 횡단보도 앞 또는 교차로에서, 차선과 수직을 이루는 선	
③	횡단보도	- 흰색 블록으로 이루어져 있으며, 블록 외에도 한 쪽 인도의 경계에서부터 다른 한 쪽 경계까지의 연장선을 횡단보도의 범위로 함	

2) 신호등/표지판

- 객체에 대한 Drawing은 내부 기준을 기반으로 정교하게 진행하며 객체 외곽 기준 최대한 타이트하게 처리
- Drawing 시 오차율은 객체 영역에 대한 부분으로 영역 오류 객체 / 전체 객체 로 계산하여 산출
- 신호등/표지판 데이터는 특성상 매우 작은 객체의 비율이 높으며, 크라우드 소싱에서 일괄적 최소크기 (예: 24x24) 가공 기준 적용이 어려운 점, 장기적 연구개발 활용성 (tiny object detection) 측면을 고려하여 일관성이 조금 떨어질 수 있더라도 가공 작업자의 판단하에 최대한 식별 가능한 가장 작은 객체까지 가공을 진행
- 신호등/표지판은 객체가 작아서 촬영 환경에 따라 객체 외곽을 명확히 파악할 수 없는 경우 (객체가 너무 작거나, 흔들림, 야간 등)가 많으며, 이때는 딥러닝 모델에서 학습이 가능하도록 외곽 드로잉에 대해 추정하여 처리

- 표지판의 내용확인은 어렵더라도 표지판임을 알 수 있을 경우, 탐지는 가능하도록 bounding box는 라벨링 하고 세부 속성은 미부여 (bbox 및 class는 필수 부여)

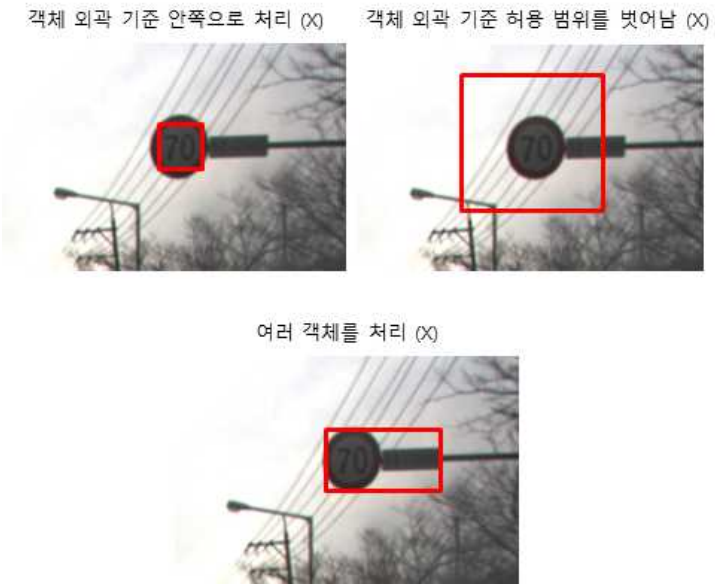
[그림 Ⅲ-116] 데이터 가공 예시

- 객체 외곽 기준 타이트하게 처리
- 촬영 환경에 따라 객체 외곽을 명확히 파악할 수 없는 경우 (작은 객체, 야간 빛번짐, 흔들림 등)에는 인공지능 알고리즘에서 디텍션이 가능한 수준으로 외곽을 추정하여 드로잉. (일반적인 물체탐지의 드로잉 기준 적용 X)

[그림 Ⅲ-117] 신호등 정상 드로잉 기준 예시

- 객체 외곽 기준 타이트하게 처리
- 신호등/표지판은 객체가 작아서 촬영 환경에 따라 객체 외곽을 명확히 파악할 수 없는 경우 (객체가 너무 작거나, 흔들림, 야간 등)가 많음. 이때는 AI 알고리즘에서 탐지가 가능한 수준으로 외곽 드로잉에 대해 추정하여 처리
- 표지판의 내용확인은 어렵더라도 표지판임을 알 수 있을 경우 탐지는 가능하도록 bounding box는 라벨링 하고 속성은 미부여

[그림 Ⅲ-118] 표지판 정상 드로잉 기준 예시



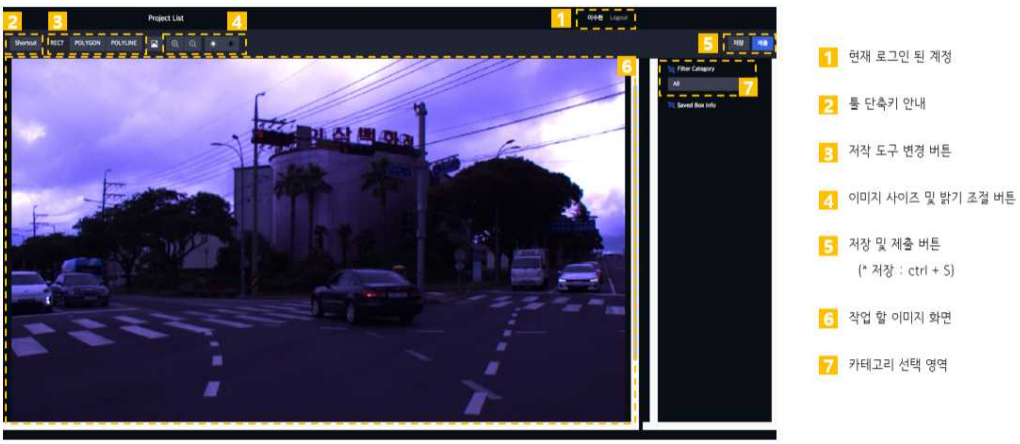
[그림 Ⅲ-119] 표지판 오류 드로잉 예시

3.3 어노테이션 / 라벨링 교육

- 신규 작업자 레이블링 교육 훈련

3.4 어노테이션 / 라벨링 도구 및 사용법

- Tool 기능 설명



[그림 Ⅲ-120] Tool 기능 설명

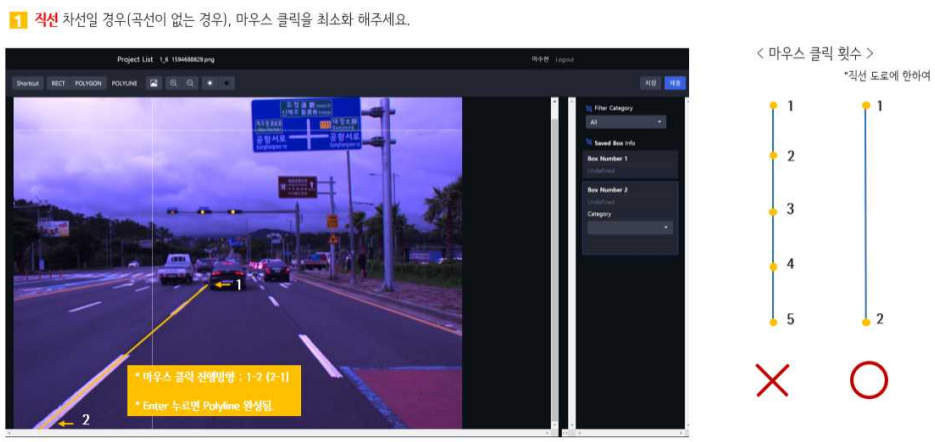
- Tool 선택 기준

[그림 Ⅲ-121] Tool 선택 기준 예시

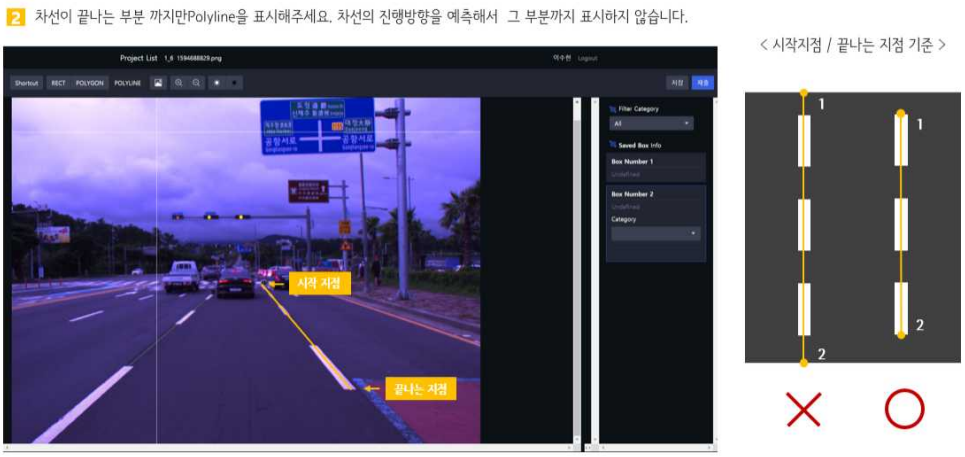
- Tool 종류와 사용법

[그림 Ⅲ-122] Tool 종류와 사용법 예시

- 객체 레이블링 유의사항



[그림 Ⅲ-123] 직선 차선일 경우 레이블링 유의사항

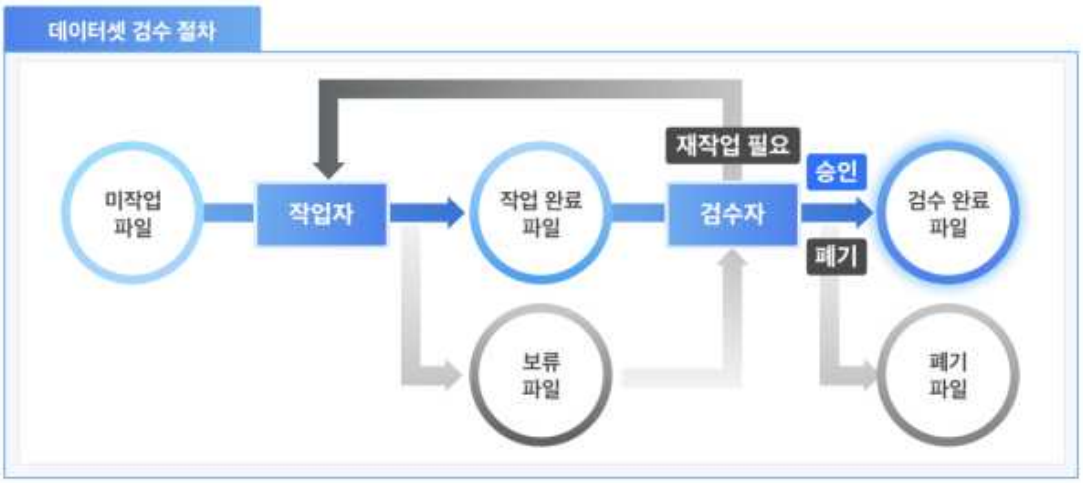


[그림 Ⅲ-124] 차선이 끝나는 부분 레이블링 유의사항

4 데이터 검수

4.1 검수 절차

1) 차선/횡단보도



[그림 Ⅲ-125] 차선/횡단보도 데이터 검수 절차

- 작업물의 상태는 대기, 작업중, 작업완료, 검수완료(1차 검수), 승인(2차 검수), 반려, 보류로 구성
- 대기: 작업물에 대해 작업자가 배정되지 않았거나, 배정되었으나 수행하지 않은 상태
- 작업중: 배정된 작업물을 수행중인 상태

- 작업완료: 작업자에 의해 작업이 완료된 상태
- 검수완료: 1차 검수자에 의해 작업물이 통과된 상태
- 승인: 2차 검수자에 의해 이중검수가 완료되고 포인트 지급이 가능한 상태
- 보류: 작업자 혹은 검수자가 판단하기 모호하거나, 작업 기준이 모호한 경우 선택하는 상태로 별도 판정단이 해당 작업물에 대해 판별한 후, 작업할 수 없다면 폐기처리, 작업할 수 있는 경우 가이드와 함께 반려 처리함
- 반려: 작업완료된 이후 프로세스에서 검수자에 의해 부적합 판정을 받은 작업물, 원 작업자에게 배정되어 재작업을 대기 중인 상태
- 개별 작업상태에서 다음 검수 프로세스 진입 후, 부적합 판정 시 판정 사유와 함께 반려처리하여 작업을 수행한 작업자에게 재배정함

2) 신호등/표지판

- 전문 검수 조직에서 단계별 절차에 따라 검수를 진행하여 99% 이상의 정확성 보장
- 목표 기준율을 넘는 오류는 가공 조직에서 재작업 진행

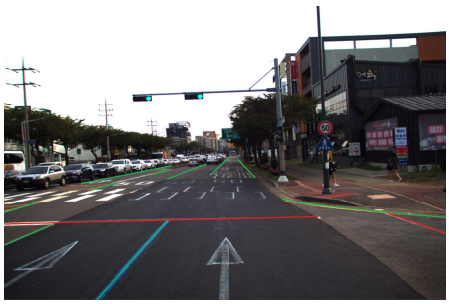
[그림 Ⅲ-126] 신호등/표지판 데이터 검수 절차

4.2 검수 기준

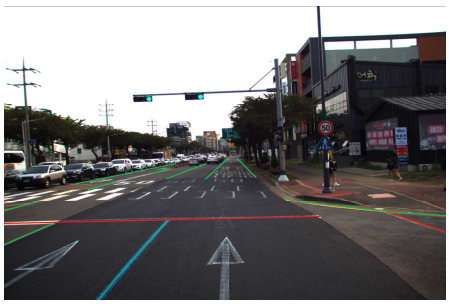
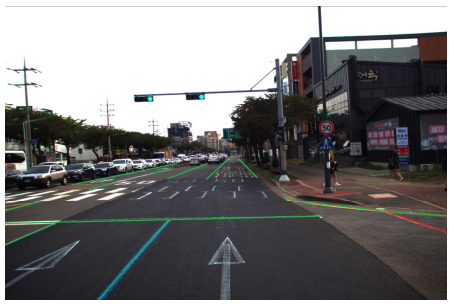
1) 차선/횡단보도

- 미검출, 과검출, 오검출 진단 및 폴리곤(폴리라인) 등의 정확도 진단 수행

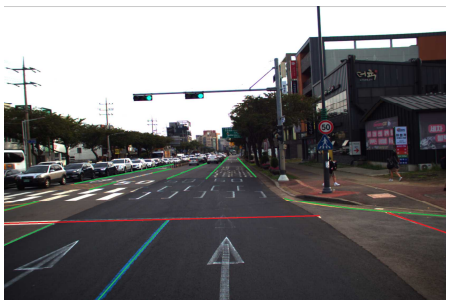
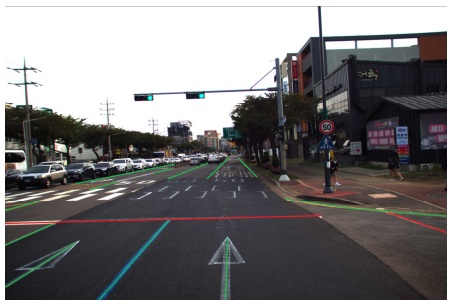
〈표 Ⅲ-109〉 반려(미검출) 예시

승인 예시	반려 예시(미검출)
	
	반려사유 : 파란 실선 차선에 대한 라벨링 미시행

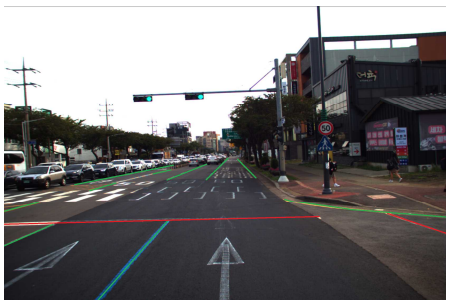
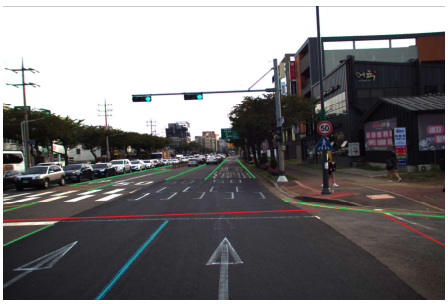
〈표 Ⅲ-110〉 반려(오검출) 예시

승인 예시	반려 예시(오검출)
	
	반려사유 : 정지선을 차선으로 라벨링

〈표 Ⅲ-111〉 반려(과검출) 예시

승인 예시	반려 예시(과검출)
	
	반려사유 : 라벨링하지 않아야 할 화살표를 차선으로 라벨링

〈표 Ⅲ-112〉 반려(정확도 부족) 예시

승인 예시	반려 예시(박스 정확도 부족)
	
	반려사유 : 정지선 라벨링의 정확도 부족

2) 신호등/표지판

- 검수 방법은 다음의 절차에 따라 진행

(TP = # of True Positive, FP = # of False Positive, FN = # of False Negative)

〈표 Ⅲ-113〉 신호등/표지판 데이터 검수 기준

진단 항목	진단 세부 항목	측정 방식	오류율 (정확도) 기준치
데이터 파일 형식 진단 (구문정확성)	• 파일 유효성 확인 – 사전 정의된 구조와 적합성 확인	#오류 파일 수 / #전체 파일 수	1% 미만
결과물 구문 정확성 진단	• JSON 파일 내부의 항목 및 형식 정확도 확인	#오류 건 (객체) 수 / # 어노테이션 파일내 전체 건 (객체)수	1% 미만
DB 구축 내용 확인 (의미정확성)	• 객체(bounding box) 탐지 정확도	• Recall = TP / (TP + FN) • Precision = TP / (TP + FP) • 정확도(F1) = 2*Recall*Precision/(Recall + Precision)	80% 이상

- 검수 기준은 다음의 목표에 따라 진행

〈표 Ⅲ-114〉 데이터 검수 기준

진단항목	세부항목	진단 방법	오류 기준
데이터 파일 형식 진단	• 파일명 정합성 확인 - 사전 정의된 구조와 적합성 확인	• Script에 의한 자동 검수 - 결과물 확장자 및 파일명 형식 확인 - 파일명의 bit 분리 후 사전 정의된 내용기준으로 확인 - 사전 정의된 형식과 적합성 확인 - 오류가 발생한 파일이름 추출 - 추출된 폴더는 사람에 의한 추가 확인	- 확장자 및 파일명 형식이 사전 정의된 내용과 일치하지 않을 경우 오류로 정의 - 사전 정의된 파일명길이를 준수하지 않을 경우 오류로 정의 - 각 bit정보가 사전 정의된 이름을 사용하지 않을 경우 오류로 정의 - 정의된 폴더구조를 지키지 않을 경우 오류로 정의
결과물 구문 정확성 진단	• 결과물 세부내용 형식 확인	• Script에 의한 자동 검수 - 결과물 내용의 형식 확인 - 세부 구성요소에 대한 확인 - 불필요 데이터에 대한 확인	- 결과물 세부 형식이 기준과 맞지 않을 경우 오류로 정의 - 결과물 요소 중 필요요소가 누락되었을 경우 오류로 정의 - 사전 정의되지 않은 요소가 존재할 경우 오류로 정의
데이터 가공 상태 진단 (의미 정확성)	• 객체 (bounding box) 탐지 정확도	• 사람에 의한 검수 - 객체 누락여부 (False Negative)에 대한 확인 - 오검출 객체 (False Positive)에 대한 확인	- 객체에 대해 bounding box가 제대로 라벨링 되었을 경우 정답으로 정의 (TP 1 추가) - 대상 객체에 대해 누락되었을 경우 오류로 정의 (FN 1 추가) - 정의되지 않은 객체에 대한 Annotation 되었을 경우 오류로 정의 (FP 1 추가)

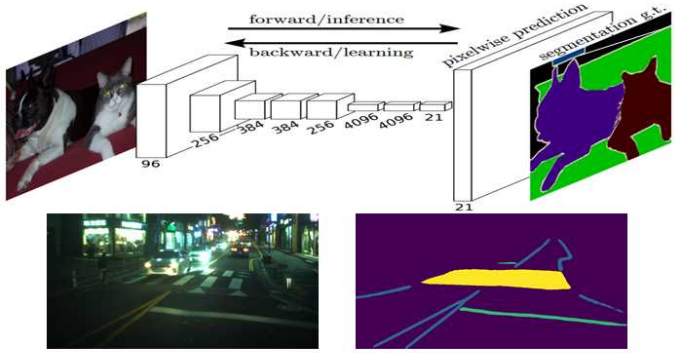
5 데이터 활용 방안

5.1 학습 모델

● 학습 모델 선정 근거

- 차선 영역 탐지 프로토타입 모델
  - 자율주행에서의 측위 알고리즘 고도화 또는 자율주행 정밀지도 제작을 위한 차선인식 AI 프로토타입 모델 공개
  - 주행 영상을 입력받아 차선 영역을 인지하는 AI 모델 개발

- 개발 환경: PyTorch, Cuda 10.1
- 학습 모델 상세 내용



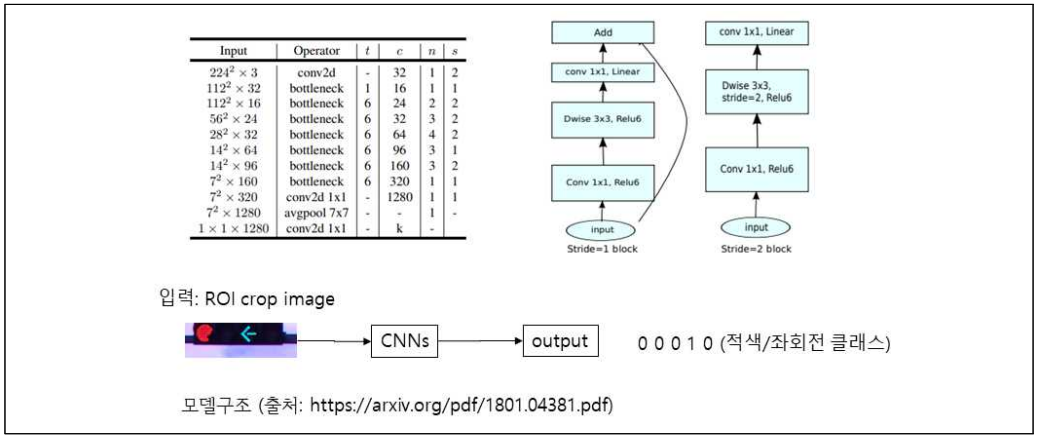
그림설명: FCN ResNet 구조 (출처: https://arxiv.org/pdf/1411.4038.pdf)

[그림 Ⅲ-127] FCN ResNet 구조를 활용한 차선/횡단보도/정지선 영역 탐지

● 학습 모델 개발 계획

1) 저연산 신호등 상태 인지 프로토타입 모델

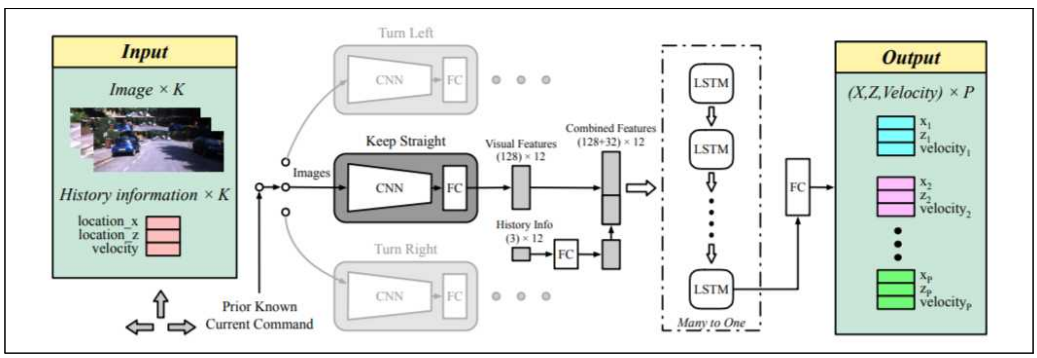
- 자율주행차량이 현재 신호등 상태에 맞는 주행계획을 세우고 안전하게 운행하기 위하여 신호등 인지 알고리즘이 요구됨
- 자율주행차는 미래에는 C-ITS와 같은 시스템으로부터 신호 정보를 수신할 수도 있으나 아직 인프라가 갖춰진 곳이 많지 않으며 이후 인프라가 구축되더라도 redundancy 측면에서도 카메라 기반의 신호 정보 인지 시스템이 필수적
- 다양한 타입의 신호등과 각각의 상태(빨간불, 노란불, 초록불 등)에 대하여 학습된 AI 모델을 신호 인지 알고리즘에 활용
- 개발 환경: PyTorch, Cuda 10.1
- 학습 모델 상세 내용



[그림 Ⅲ-128] MobileNetV2 기반의 구조를 활용한 신호등 상태 인지

2) 영상, GNSS/INS 데이터 기반의 차량궤적 예측 모델

- 자율주행에서의 차량궤적 예측 및 추천을 위한 AI 모델 공개
- 주행 영상, 측위 결과 및 속도 데이터 시퀀스를 입력받아 차량궤적을 추측하는 AI 모델 개발
- 구축된 다중센서 원시데이터를 활용하며, 차선/횡단보도/신호등/표지판 등의 라벨링 데이터 뿐아니라 다양한 주행 데이터의 활용성 발굴을 위한 프로토타입 모델
- 개발 환경: PyTorch, Cuda 10.2
- 학습 모델 상세 내용



[그림 Ⅲ-129] CNN-LSTM+State 네트워크 기반의 모방 학습을 통한 궤적 예측 모델

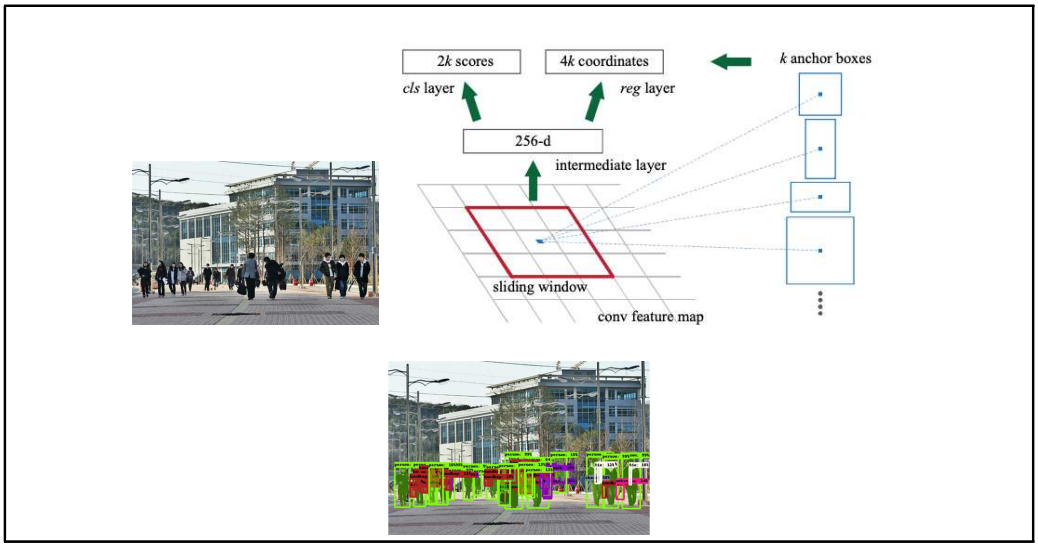
- 서비스 활용 시나리오: 궤적 예측에 필요한 입력값의 간소화를 통한 보다 범용적이고 제한이 적은 차량 궤적 예측

〈표 Ⅲ-115〉 활용 모델과 제안모델의 차량 궤적 예측에 필요한 입력값 비교

	영상, LiDAR 활용 모델	제안된 모델
필요 입력값	• 영상 데이터 • LiDAR 스캔 데이터 • GPS 데이터 • 차선 데이터 • 정적 객체 데이터 • 이외의 모델별 필요 데이터	• 영상 데이터 • GPS 데이터 또는 INS 데이터

3) 보행자 위치 추정 알고리즘

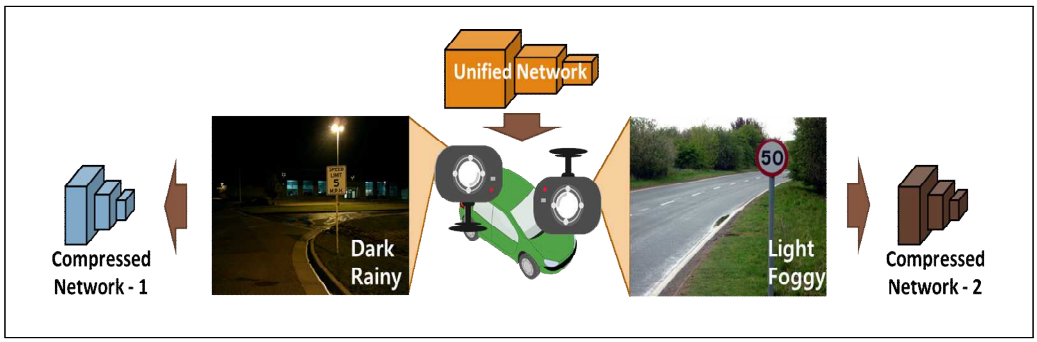
- 보도/횡단보도 근방의 보행자 위치 추정과 예측은 안전한 자율주행을 위한 중요한 선행 조건임. 특히 횡단보도가 아닌 곳에서 차도를 건너는 보행자는 빠르고 정확한 검출이 사고 예방에 필수적임. 보행자의 경우 짧은 영상에서는 정적으로 보일 수도 있는데, 정적, 동적 보행자 두 경우 모두를 놓치지 않고 검출해야 함
- 자율주행에서의 보행자 위치 추적을 위한 보행자 검출 AI 프로토타입 모델 공개
- NIA AI 데이터 구축사업의 타과제(동적 객체 인지)와의 연계 활용 기대
- RGB 기반의 주행 영상을 입력받아서 보행자 bounding box를 인지하는 AI 모델 개발
- 개발 환경: Cuda 10.1, PyTorch, TensorFlow
- 모델 상세 내용



[그림 Ⅲ-130] Faster-RCNN 기반의 보행자 영역 탐지

- ResNet-50 기반의 Feature extraction을 전체 이미지에 대해 수행
- ROI Pooling을 수행하여 임의의 size의 bounding box에 대해서 feature extraction을 수행
- 각각의 bounding box에서 얻어진 feature vector에서 보행자이면 보행자임을 알려줌. 보행자의 위치 특성 감지
- 서비스 활용 시나리오: 해당 모델이 고도화될 경우 레이블링 작업에 큰 도움을 줄 수 있음
- 중복적인 탐지를 통해 보행자의 영역을 보수적인 합집합으로 고려하여 자율주행 안전성을 높일 수 있음

4) 효율 표지판 탐지를 위한 자율주행 환경 적응형 네트워크 모델 압축 기술



[그림 Ⅲ-131] 효율 표지판 탐지를 위한 자율주행 환경 적응형 네트워크 모델

- '환경 적응형 네트워크 모델 압축 기술'은 자율주행환경에서 사용되는 인공지능 네트워크의 성능을 유지하면서 연산에 필요한 메모리와 계산을 줄이기 위한 AI 시범 서비스 모델
- 본 시범 서비스 모델은 단순하게 인공지능 네트워크를 압축하는 것이 아니라 현재 입력되고 있는 환경에 적합한 압축 방식으로 자동으로 찾아 선택하는 방법
- 기존 인공지능 네트워크 압축 기술보다 더 높은 압축율과 성능 유지율 기대 가능
- 특히 이번 과제를 통해 확보되는 데이터셋의 경우 1개 특정 도시가 아닌 매우 다양한 도시에서 데이터셋을 확보하기 때문에 다양한 환경에 적응하여 자동으로 인공지능 네트워크를 압축하는 기술을 검증하기에 적절
-

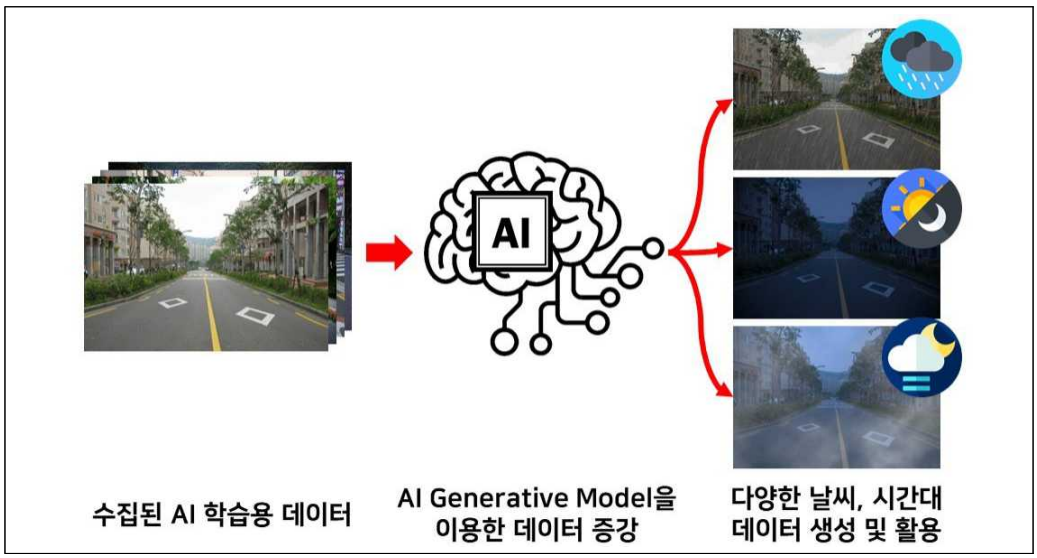
- 프로토타입 모델은 표지판이 촬영된 칼라 영상(이미지)를 입력으로 하고 표지판의 종류를 구별하는 객체분류 네트워크이며, 이 때 일관된 환경에 해당 네트워크를 지속적으로 활용하면, 네트워크의 메모리와 계산량이 그 환경에 최적화하여 점점 줄어들게 됨
- 단, 이 때 기존의 성능은 그대로 유지할 수 있도록 설계되어 있음

● 환경 적응형 네트워크 모델 압축 기술

〈표 Ⅲ-116〉 환경 적응형 네트워크 모델 압축기술 활용 방안 세부 내용

구분	세부 내용
딥러닝 연산 가능한 고성능 GPU 활용	• 입력되는 표지판의 종류를 자동으로 인식 • 1개 GPU에 대해 많은 인식 알고리즘을 구동할 수 있도록 각 네트워크가 점점 적응적으로 압축됨
원천 데이터 확보 및 활용 방안	• 서로 다른 환경에서 촬영된 영상들에서 표지판을 추출하여, 각 환경에 적응된 압축 네트워크를 확인 • 기존 단순 압축 기술에 비해 환경 적응형 알고리즘이 더 우월한 성능을 보임을 확인할 수 있음
모델 구현 방안	• 확률적 네트워크 블록 제거 (Stochastic Block-out) 알고리즘을 활용하여 네트워크를 학습 • 서로 다른 두 개 네트워크의 파라미터를 공유 (Weight sharing) 할 수 있도록 네트워크를 학습 • 이후 실제 새로운 환경에서 구동될 때 블록제거와 파라미터 공유를 선택적으로 활용하여, 현재 환경에 최적의 모델을 결정할 수 있도록 설계

5) 생성모델 기반 데이터 증강을 이용한 신호 현시정보 인식 AI 모델 고도화



[그림 Ⅲ-132] AI Generative Model 기반의 현시정보 인식 AI 모델

- '생성모델 기반 데이터 증강 기술'은 자율주행환경에서 희귀하게 촬영 가능한 이미지에 대해 높은 인공지능 인식 성능을 얻을 수 있도록 하는 시범 서비스 모델
- 일반적인 자율주행 데이터셋을 인공지능 학습에 활용할 경우 악천후나 짙은 안개가 존재하는 영상 등은 일반적인 맑은 날씨에 촬영된 영상보다 그 수가 훨씬 적기 때문에, 일반적인 환경보다 특수한 환경에서의 성능이 크게 떨어진다는 단점이 있음
- 하지만 이런 특수한 환경에서의 데이터를 실제로 일반적인 환경 데이터 개수만큼 얻는 것은 불가능하며, 이런 문제를 해결하기 위해 일반적인 환경에서 촬영된 영상을 특수한 환경에서 촬영된 것처럼 보이도록 변환하는 생성모델을 생성
- 이를 통해 일반적인 환경의 영상을 특수한 환경으로 변환할 수 있고, 기존보다 훨씬 많은 개수의 특수한 환경 데이터를 확보할 수 있게 됨
- 특히 본 과제에서 취득하고자 하는 데이터에는 악천후 등을 포함한 특수한 환경 데이터를 다수 포함하기 때문에 이렇게 개발된 생성모델의 효율성을 측정하기에 적합
- 프로토타입 모델은 신호등이 촬영된 칼라 영상(이미지)를 입력으로 하고 신호등의 위치를 탐지하는 객체인식 네트워크
- 기존 인식 모델의 경우 데이터가 많이 제공되는 일반적인 환경에서는 높은 성능을 보이지만 특수한 환경에서는 데이터 부족으로 인해 성능이 크게 저하되지만, 본 기술을 통해 확보한 추가적인 특수한 환경 영상을 학습에 추가로 활용하여 모든 상황에 일반적으로 적용 가능한 인공지능 모델을 확보 가능

● 생성모델 기반 데이터 증강 기술

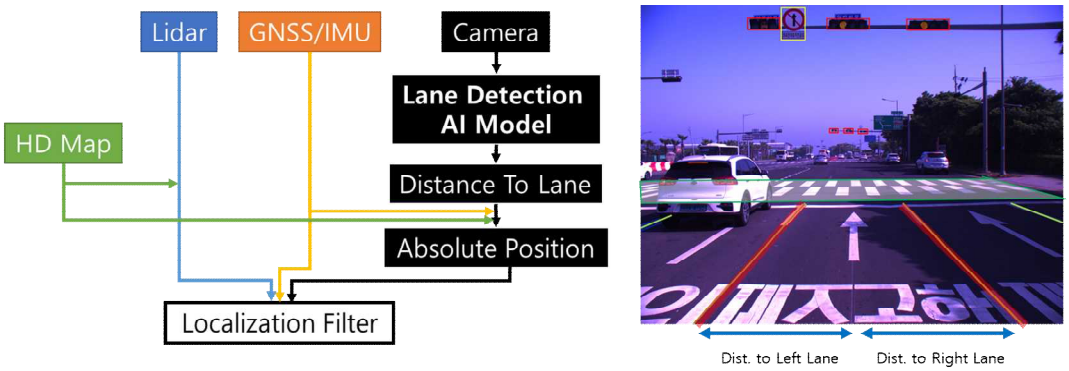
〈표 Ⅲ-117〉 생성모델 기반 데이터 증강기술 세부 내용

구분	세부 내용
딥러닝 연산 가능한 고성능 GPU 활용	• 입력되는 일반적인 영상을 특수한 환경으로 변환 • 생성 모델의 경우 실시간 구동이 필요없기 때문에 다수의 GPU를 활용해 최대한 많은 개수의 데이터를 확보할 수 있도록 설계 및 구성
원천 데이터 확보 및 활용 방안	• 악천후 등 특수한 상황에 촬영된 영상을 따로 분류하여 일반적인 상황과 특수한 상황을 독립적으로 데이터셋을 구성 • 이후 여러 일반적인 상황 데이터를 특수한 상황처럼 보이도록 학습 진행

구분	세부 내용
모델 구현 방안	• 영상 환경 변환 기술을 활용하여 입력된 영상의 중요한 정보는 유지하면서 촬영된 환경만 바꿀 수 있도록 생성 모델 학습 • 이렇게 얻은 데이터들을 개별적으로 저장할 경우 데이터의 크기가 무한대로 늘어나기 때문에, 모든 데이터를 활용하기 보다는 실제로 학습에 도움이 될 수 있는 데이터를 선택적으로 저장 • 이후 완전히 새롭게 촬영된 데이터셋이 입력되더라도 충분히 학습된 생성모델을 활용하면 새로운 데이터셋에 대해서도 특수한 환경의 추가 영상을 확보할 수 있음

5.2 서비스 활용 시나리오

- 차선인식 AI 모델의 자율주행 활용 예시는 아래 그림과 같음



[그림 Ⅲ-133] 차선인식 AI모델의 자율주행 측위 알고리즘에서의 활용 예시

제12장 시설작물 질병진단 이미지 데이터

1 데이터 정보 요약

1.1 가이드 분류

대분류	이미지	중분류	비전	소분류	JPEG

1.2 데이터 정보

데이터 이름	시설작물 질병진단 이미지 데이터
데이터 요약	시설작물에 발생하는 주요 병해에 대해 발병 단계별 이미지 수집
데이터 출처	전국 시설작물 실 재배 온실 / 각 대학연구소 실험 온실

1.3 데이터 구축 개요

- 농산물 이미지 학습용 데이터 구축을 위한 객체 및 속성정보를 취득하여 정보이용자(농민, 일반인)가 연구개발에 쉽고 효율적으로 활용할 수 있는 데이터를 제작하고자 함. 이를 위해 인공지능 학습을 위한 학습용 시설 질병 이미지 데이터를 수집, 정제를 거친 다음 메타데이터가 추가된 시설 작물의 질병을 판단하는 AI 학습용 자료를 만드는 것이 목적임

[그림 Ⅲ-134] 학습용 데이터 구축 공정도

1.4 구축 목적

- 시설 재배작물 12종의 고추마일드모틀바이러스병 등 20종의 식물병에 대한영상 데이터 수집 및 가공을 통해 고품질의 학습 데이터를 구축하기 위해 작성되었음
- 학습된 인공지능은 농가 내 설치된 고해상도 카메라를 통해 수집된 각종 작물의 상태를 분석하고, 분석 결과를 기반으로 병해충 예찰 및 진단을 수행함
- 이를 통해 농가의 예찰 노동력을 절감하고, 상황 발생 시 즉각적인 방제를 수행하여 작물 생산성을 보호하고, 이를 통해 농가 경제 경쟁력을 향상시킬 수 있음
- 아울러 본 사업을 통해 구축된 AI 학습용 빅데이터와 AI 진단 모델은 공공데이터로 공개되어, 민, 관, 학 연계를 통한 정밀 방제 기술 개발에 기여하고, 스마트 농업 분야 국가 경쟁력 향상에 이바지하기 위함

1.5 활용 분야

- 이미지를 통해 시설작물 발생 병해 진단

1.6 유의 사항

1) 데이터 수집 규모의 미달 가능성 및 해결 방안

- 제한 된 연구기간으로 인하여 작물별 생육기에 따른 병해충 및 정상작물의 데이터 수집에 한계가 있다. 지역별 각기 다른 작기를 조사하고, 다양한 작부체계를 사전 조사하여 전 생육기에 대한 데이터 확보에 주력
- 초발 상의 작물병의 진단을 현장에서 단순 외관상의 증상만으로 동정하여 데이터를 확보할 경우 오류가 있을 수 있다. 따라서 전문가 자문을 통하여 해당 병해로 판정된 이미지만을 확보
- 과제 종결 시점에서 정제 기준 미달 데이터, 불량 및 중복 데이터로 인한 AI 학습데이터의 저품질 발생에 의해 수집 교모가 목표에 도달 하지 못할 경우가 발생할 수 있다. 따라서 고해상도(1Mb 이상) 이미지를 수집하여 저해상도 데이터 불량을 사전에 차단함. 해당 병해 이미지 수집목표인 1000장의 50% 이상 여분의 고해상도(1Mb 이상) 이미지를 수집
- 정상작물의 데이터는 작물의 품종, 재배환경에 따라 엽의 형태, 과의 형태가 차이가 있어 단일품종 혹은 환경에 데이터를 해석의 오류가 발생한다. 계획단계에서 이를 구분하고, 이에 대응할 수 있는 데이터를 확보
- 목적으로 하지 않은 병해충에 의하여, 대상작물이 피해를 입어 독립적 증상을 해석하기 어려운 경우 발생 할 수 있다. 대상 병해 이외의 비 목적 병해충은 선택적 농약을 선택하여 방제하여 목적 병해충만을 유지하여, 작물 부위별, 감염 정도별, 시기별 다양한 피해 증상에 대한 이미지 데이터를 확보

2) 유사 병징으로 인한 데이터 획득의 해결방안

- 전년도 다발생하였던 포장을 선정하여, 대상 질병 발생을 구축
- 시설작물은 잡초발생 문제로부터 자유스러우므로 제초제 피해에 의한 유사 병징 데이터 오류는 발생하지 않지만 생리장애의 경우 바이러스병과 유사하여 유사 병징 데이터 오류가 발생할 수 있는데 바이러스병은 전염성이지만 생리장애는 비전염성 이라는 특징을 가지고 있는 점을 이용하여 바이러스병에 의한 병해 이미지만을 획득함. 무기양분 결핍에 의한 바이러스 유사증상이 나타난 경우 엽면시비를 통하여 생리장애가 회복 가능하므로 엽면시비법을 통하여 판별하여 바이러스병에 의한 병해 이미지만을 획득하는 것도 또한 해결 방안임

3) (개인 정보) 농장에서 촬영 시, 세부 지번이나 농장주의 인적사항이 포함되지 않도록 정제

4) (저작권 및 활용범위) 농가 방문촬영시 사전 저작권 및 AI Hub 업로드 등이 공공활용이 가능한 내용이 포함된 촬영동의서 기명 후 서명날인 청구

5) (데이터 가공 및 배포) 모든 데이터는 2차 가공 및 배포 금지, 사용 시 출처 표기 관련 연구

2 데이터 획득 및 정제

2.1 원시 데이터 선정

1) 고추마일드모틀바이러스병, Pepper mild mottle virus

- 시설 재배 고추에서 고추마일드모틀바이러스병에 대하여 병징의 영상 데이터를 수집하고 각 코드 부여
- 작물: 고추
- 질병: 고추마일드모틀바이러스병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 전북 전주시

2) 고추점무늬병, Xanthomonas campestris

- 시설재배 고추에 발생하는 고추 점무늬병에 대하여 병징 및 표징의 영상 데이터를 수집하고 각 작물 및 질병별로 코드 부여
- 작물/병해 종류: 고추 / 고추 점무늬병
- 생육단계: 대상 병해별 발병초기

3) 토마토황화잎말이바이러스병, Tomato yellow leaf curl virus

- 시설 재배 토마토에서 토마토황화잎말림(잎말이)바이러스병에 대하여 병징의 영상 데이터를 수집하고 각 코드 부여
- 작물: 토마토
- 질병: 토마토황화잎말이바이러스병

- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 경북 안동시

4) 토마토잎곰팡이병, Fulvia fulva
- 시설재배 토마토에서 토마토잎곰팡이병에 대하여 병징의 영상 데이터를 수집하고 각 코드 부여
- 작물: 토마토
- 질병: 토마토잎곰팡이병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 대전광역시 유성구/충남 예산군

5) 오이녹반모자이크바이러스, Cucumber mosaic virus
- 시설 재배 오이에서 오이녹반모자이크바이러스병에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여
- 작물: 오이
- 질병: 오이녹반모자이크바이러스병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 서울특별시 관악구

6) 포도노균병, Plasmopara viticola
- 시설 재배 포도에서 포도노균병에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여
- 작물: 포도
- 질병: 포도노균병
- 질병정도 혹은 단계: 발병 초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 전북 김제시/정읍시

7) 수박탄저병, Colletotrichum orbiculare
- 시설 재배 수박에서 수박 탄저병에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여
- 작물: 수박

- 질병: 수박 탄저병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 전북 고창군

8) 수박흰가루병, Sphaerotheca fusca

- 시설 재배 수박에서 수박 흰가병에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여
- 작물: 수박
- 질병: 수박 흰가루병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 전북 고창군

9) 참외흰가루병, Sphaerotheca fusca

- 시설 재배 참외에서 참외 흰가루병에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여
- 작물: 참외
- 질병: 참외 흰가루병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 경북 성주시/안동시

10) 참외노균병, Pseudoperonospora cubensis

- 시설 재배 참외에서 참외 노균병에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여
- 작물: 참외
- 질병: 참외 노균병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 경북

11) 딸기흰가루병, Spharotheca humuli

- 시설재배 딸기에서 딸기흰가루병에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여

- 작물: 딸기
- 질병: 딸기흰가루병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 충남 논산시/천안시

12) 딸기잿빛곰팡이병, Botrytis cinerea

- 시설재배 딸기에서 딸기잿빛곰팡이병에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여
- 작물: 딸기
- 질병: 딸기잿빛곰팡이병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 충남 논산시/천안시

13) 가지흰가루병, Golovinomyces cichoracearum

- 시설 재배 가지에서 가지 흰가루병에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여
- 작물: 가지
- 질병: 가지 흰가루병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 경기도 여주시/경북 영천시

14) 가지잎곰팡이병, Mycovellosiella nattrassii

- 시설 재배 가지에서 가지 잎곰팡이병에 대하여 병징의 영상 데이터를 수집하고 각 코드 부여
- 작물: 가지
- 질병: 가지 잎곰팡이병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 경기도 여주시/경북 영천시

15) 상추균핵병, Sclerotinia minor

- 시설 재배 상추에서 상추균핵병에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여

- 작물: 상추
- 질병: 상추균핵병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 광주광역시 광산구

16) 상추노균병, Bremia lactucae

- 시설 재배 상추에서 상추노균병에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여
- 작물: 상추
- 질병: 상추노균병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 광주광역시 광산구

17) 단호박흰가루병,Sphaerotheca fuliginea

- 시설 재배 단호박에서 단호박 흰가루병에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여
- 작물: 단호박
- 질병: 단호박 흰가루병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 강원도 춘천시

18) 단호박점무늬병, Pseudomonas syringae pv. syringae

- 시설 재배 단호박에서 단호박 점무늬병에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여
- 작물: 단호박
- 질병: 단호박 점무늬병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 강원도 춘천시

19) 오이녹반모자이크바이러스, Cucumber green mottle mosaic virus

- 시설 재배 쥬키니호박에서 오이녹반모자이크바이러스에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여
- 작물: 쥬키니호박
- 질병: 오이녹반모자이크바이러스
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 서울특별시 노원구

20) 애호박점무늬병, Pseudomonas syringae

- 시설 재배 애호박에서 애호박점무늬병에 대하여 병징의 영상 데이터를 수집하고 각 코드를 부여
- 작물: 애호박
- 질병: 애호박점무늬병
- 질병정도 혹은 단계: 발병초기, 중기, 말기의 단계별 식별 가능한 수준의 데이터 확보
- 획득지역: 충북 청주시

2.2 규제관련 사항

- 촬영물에 대한 저작권(사용 허락 동의 여부) 확인, 사용 허락 동의에 대한 법적 고지 의무 준수
- 획득시, 개인정보 노출 부분 비식별화
- 획득시, 저작권 비동의 이미지나 이미지 자체가 개인정보인 경우는 비식별화

※ AI 학습시 포함되었던 온도와 습도 범위를 벗어난 상태에서 작물에 병이 발생하면 AI를 이용하여 병징 판독시 오류 발생 가능성이 존재함. 따라서 환경센서 데이터(온도/습도, 토양정보센서, 양액정보센서, 외부기상)입력은 고려하지 않음

※ 이미지를 획득한 위치에 관한 요구도 가 있을 수 있으나 많은 이미지가 디지털카메라를 이용하여 획득하거나 인터넷 연결이 되지 않는 상황에서 스마트폰카메라를 이용하여 획득하므로 위치정보 입력은 고려하지 않음

※ 일정 시간대에만 이미지 획득을 하는 것이 아니라 이른 오전부터 늦은 오후의 다양한 시각대에 작물 생육기간 동안 이미지 데이터를 획득할 경우 다양한 조도대에 노출이 이루어 진다. 다시 말해 특정 조도대에서 획득한 이미지를 이용하여 발생할 수 있는 Overfitting문제가 발생하지 않을 것으로 보이므로 조도를 변화시키면서 사진 촬영하는 것은 고려하지 않음

※ 작물에 발생한 병해 이미지는 정적인 대상으로 가까운 거리에서 초점을 맞추워 병해 및 병징을 촬영하여 AI 학습을 위한 뚜렷한 병해 이미지를 획득하는 것이 목적인만큼 이미지 획득을 위한 거리를 특정하는 것은 고려하지 않음

2.3 획득 및 정제 절차

1) 조사표본 : 병이 자연 발생한 작물 및 대상 작물병을 유발하는 병원체를 접종하여 병을 유도하여 병해가 나타나는 작물을 이미지 획득 표본으로 선정(포도 노균병의 경우 포도농장을 직접 방문하여 포도노균병 발생을 확인하고 표본으로 선정하여 병 진전도에 따라 이병엽의 윗면과 뒷면을 촬영)

- 온실 내 환경편차를 줄이도록 온실 중앙에 위치시켜 재배관리를 진행한 식물체를 조사 대상으로 함(식물재배실에서 포트에서 재배할 수도 있음)
- 토마토잎곰팡이병의 경우 자연발생한 포장에서 촬영. 병발생 포장 입구에서부터 안쪽까지 전체를 훑어가며 사진 촬영 데이터 확보
- 딸기흰가루병의 경우 병원균을 인공접종 및 자연발병 한 식물체에서 사진 촬영 데이터 확보
- 딸기잿빛곰팡이병의 경우 병원균을 인공접종 및 자연발병 한 식물체에서 사진 촬영 데이터 확보

2) 질병 발생 및 피해증상

- 대상 시기는 전생육기에 걸쳐 작물에 발생되는 병해의 이미지를 시기별(발병초기, 발병중기, 발병후기), 부위별(잎, 가지, 과실)로 병징을 구분하여 촬영함
- 가능한한 대상 병해의 발병초기, 발병중기, 발병후기 이미지 획득 건수가 균일하게 포함되도록 획득함
- 실험실에 접종관리하는 식물을 이용하여 야간에도 상황 재현하여 촬영 가능

3) 수집장비

- 디지털카메라/스마트폰 단말기/스마트폰 단말기(전용App설치) /자동수집 장비
  ① 디지털카메라: 100만 화소 이상 이미지 촬영 가능한 카메라
  ② 스마트폰 단말기(전용App설치 스마트폰 포함): 100만 화소 이상 이미지 촬영 가능 단말기 전체
  ※ 휴대 편리한 장비로 불특정 환경에서 촬영 가능 // 수집앱을 통한 효율적인 업로드 가능
- 파일형식 : JPEG 또는 PNG

4) 중복 이미지 회피: AI 다양한 형태의 작물병 이미지가 수집될 수 있도록 하며, 만일 동일한 이미지 대상을 수집하고자 할 때는 일주일 간격을 두고 이미지를 수집함(병진전이 느린 경우 중복이미지로 판정이 날 수 있는 경우를 피하기 위함)

5) 각도: 데이터 신뢰도를 높이기 위해 이미지 획득시 촬영 각도를 특정하지 않고, 병징의 특성을 확인할 수 있는 각도로 촬영

6) 대상: 병징 전체가 나오도록 초점이 정확하게 맞아야 하며, 이미지 촬영 거리에 관계없이 대상을 명확하게 보여주도록 촬영

7) 조도: Overfitting을 방지하기 위해 일정 시간대에만 이미지 획득을 하는 것이 아니라 이른 오전부터 늦은 오후의 다양한 시각이대에 생육기간 동안 수개월에 걸쳐 대상 병징을 촬영하므로 조도 변화를 주면서 촬영하지 않으나 실험실에서 접종관리하는 작물병징 촬영 시 Overfitting을 방지하기 조도 변화를 주면서 촬영

8) 해상도: 고해상도(100만화소 이상)

9) 촬영이미지 및 정제이미지 쌍을 보존하여, 촬영장비를 통해 기록된 메타데이터(exif 정보가)가 보존될 수 있도록 함

10) 이미지 획득 환경 정보

- 경우에 따라 획득/촬영 당시의 주변환경 정보(온/습도, 조도, 날씨 등)를 획득 및 관리
- 날씨 단위는 맑음, 구름, 흐림, 비, 눈 5종으로 한정

11) 정제 방법

- 초점이 잘못되었거나 중복에 해당하는 이미지는 제거 후 여분의 다른 이미지로 대체함
- 도구 등으로 데이터 학습에 적합하도록 노이즈 제거 및 엣지 강화
- 개인정보 보호등의 사유로 비식별화가 필요하지만 획득단계에서 비식별화가 되지 않은 경우 비식별화 추가

2.4 획득 및 정제 기준

1) 고추마일드모틀바이러스병, Pepper mild mottle virus

- 고추마일드모틀바이러스병은 전생육기에 걸쳐 잎, 가지, 과실에 발생하나, 잎에 병징이 먼저 나타나고 쉬게 확인이 가능하므로 잎의 병해 이미지를 획득함

- 잎에 나타나는 고추마일드모틀바이러스병은 ❶잎색이 균일하지 않은 모틀증상과 ❷요철 증상이 나타나면서 말리는 두 가지 형태로 나타남
- 모틀증상이 나타난 후 병이 심해지면 말리는 경향이 있어, 모틀증상은 초기 피해로 간주함
- 모틀증상이 나타난 후 잎말림이 보이면 중기 피해로 간주함
- 모틀증상 없이 요철증상이 나오기 시작하는 경우에도 초기 피해로 간주함
- 요철증상이 명확하게 나오고 잎말림이 확실해지면 중기 피해로 간주함
- 잎말림이 보이면서 뒷면이 확실하게 보이면 말기 피해로 간주함

2) 고추점무늬병, Xanthomonas campestris

- 고추 점무늬병은 세균에 의한 병으로 잎에 처음에는 회갈색의 작은 점무늬로 나타나고, 진전되면 중심부는 흰색으로 변하며 병반의 가장자리는 암갈색을 띰. 병반의 주위에는 황색의 테두리가 형성됨. 신함 경우에는 잎 전체가 갈색으로 변해 떨어짐

3) 토마토황화잎말이바이러스병, Tomato yellow leaf curl virus

- 토마토황화잎말이바이러스병은 전생육기에 걸쳐 잎, 가지에서 주로 발생하나, 잎에 병징이 먼저 나타나고 쉽게 확인이 가능하므로 잎의 병해 이미지를 획득함
- 잎에 나타나는 토마토황화잎말이바이러스병은 정상잎에 비해 잎 크기가 작아지고 곱슬거리는 것처럼 변형(curling)되며 노란색으로 잎의 색이 변하는 병징으로 나타남
- 정상 잎에 비해 잎이 말리는 경향이 보이면, 황화잎말이병의 초기 피해로 간주함
- 잎의 크기가 전반적으로 위축이 되며 말리고, 잎 가운데 부위가 돌기가 형성되는 것처럼 곱슬거리기 시작하면 중기 피해로 간주함
- 잎말림과 동시에 잎의 색 일부가 노란색으로 변화하면 말기 피해로 간주함

4) 토마토잎곰팡이병, Fulvia fulva

- 전생육기에 걸쳐 작물의 잎에 발생되는 병해이나 보통은 살균제를 이용한 방제작업의 실시로 재배 중 자연발병을 관찰하기 어려움
- 농가 수확 완료로 살균제 처리가 이루어지지 않은 포장에서 2주간에 걸쳐 촬영한 후 발병 정도에 따라 시기별(발병 초기, 발병 중기, 발병 말기)로 분류
- 잎의 표면에 담황색의 작은 반점이 발생하기 시작 하면 초기 피해로 간주함

- 잎의 표면에 담황색의 작은 반점 증가하면서 병반이 확대 되면 중기 피해로 간주함
- 잎은 연한 녹색이나 갈색으로 썩고 오그라들면서 말라죽는데 말기 피해로 간주함

5) 오이녹반모자이크바이러스, Cucumber mosaic virus

- 오이녹반모자이크바이러스병은 전생육기에 걸쳐 잎, 과실에 발생하며 잎에 병징이 먼저 발생하는 이미지를 획득함
- 잎에 나타나는 오이녹반모자이크바이러스병은 연녹색의 반점이 보이는 모자이크 증상으로 나타남
- 연녹색 반점이 나타나면 초기증상으로 간주함
- 연녹색 반점이 증가하고 잎 전체적으로 발생하면 중기 피해로 간주함
- 연녹색의 반점 발생이 진전되어 황화 증상이 나타나면 말기 피해로 간주함

6) 포도노균병, Plasmopara viticola

- 포도노균병은 전생육기에 걸쳐 잎과 과실에 발생하나, 대부분 잎에 병징이 뚜렷하고 과실에 병징이 매우 드믈기 때문에 잎의 병해 이미지를 획득함
- 잎에 나타나는 포도노균병은 ❶잎 윗면에는 모무늬 증상을 일으키고, ❷잎의 뒷면에는 하얀 서릿발 같은 곰팡이균체가 형성됨
- 발병 초기에 잎에는 연두색 대니 담황색의 병반이 형성되며, 병의 진전과정에서 엽맥을 넘지 못하고 막히는 현상으로 인해 병반은 모무늬와 유사함
- 발병 후기에는 병반 및 잎 전체가 갈색으로 변하고 잎말림이 보이며 조기 낙엽이 발생함
- 잎 뒷면에는 다각형 부분에 하얀 서릿발 같은 곰팡이균체(포자낭경 및 포자낭)가 발생함
- 잎 윗면에 모무늬 증상이 나타나고, 잎 아랫면에 곰팡이균체가 형성되면 초기 피해로 간주
- 잎 윗면에 모무늬 병반이 병합되고, 아랫면에 곰팡이균체가 대량 형성되면 중기 피해로 간주
- 잎 말림이 보이면서 뒷면이 곰팡이균체로 덥히면 말기 피해로 간주함

7) 수박탄저병, Colletotrichum orbiculare

- 수박탄저병은 잎, 잎자루, 줄기, 과실 및 과경에 발생한다. 잎에서는 처음에 갈색의 부정형 반점으로 나타나고, 진전되면 암갈색의 겹무늬 증상으로 확대되므로 잎에서 먼저 쉽게 확인이 가능하므로 잎의 병해 이미지를 획득함

- 잎에 나타나는 수박탄저병은 ❶잎에 갈색의 부정형 반점과 ❷심하게 감염된 잎은 병반이 융합하여 커지면서 회흑색으로 변하는 형태로 나타남
- 잎에 수침상의 황색 병반이 보이기 시작하면 초기 피해로 간주함
- 잎에 병반이 많이 퍼지고 병반이 융합하여 잎이 회흑색으로 변하면 중기 피해로 간주함
- 감염부위가 말라서 부서지기 시작한 것을 말기 피해로 간주함

8) 수박흰가루병, Sphaerotheca fusca

- 수박 흰가루병은 잎과 줄기에 발생하며 처음에는 흰색의 균총이 불규칙한 원형으로 나타나고, 진전되면 잎전체가 밀가루를 뿌려 놓은 것 같은 증상을 가진 잎의 병해 이미지를 획득함
- 잎에 어린 식물조직 위에 밀가루를 뿌려놓은 듯이 크고 작은 흰색의 병반을 형성하면 초기 피해로 간주함
- 잎에 병반이 많이 퍼지고 잎 전체에 백색 또는 회백색의 병반을 덮으면 중기 피해로 간주함
- 감염부위가 잎이나 다른 기관이 완전히 흰가루에 뒤덮인 듯한 병징을 말기피해로 간주함

9) 참외흰가루병, Sphaerotheca fusca

- 참외 흰가루병은 전생육기에 걸쳐 잎, 줄기 과실에 발생하며, 주로 잎에 발생하는 병징은 쉽게 확인이 가능하므로 잎의 병해 이미지를 획득함
- 잎에 발생하는 참외 흰가루병은 흰색의 균총이 작은 원형의 형태로 분산되어 나타남
- 잎의 부분적으로 작은 원형의 흰색 균총이 보이면 초기 피해로 간주함
- 잎에 발생한 흰색의 균총이 잎의 절반 이상 관찰되면 중기 피해로 간주함
- 잎 전체에 흰색의 균총이 뒤덮게 되면 말기 피해로 간주함

10) 참외노균병, Pseudoperonospora cubensis

- 참외 노균병은 생육 중기 이후부터 잎에 발생하며, 엷은 황색의 병반을 쉽게 확인이 가능하므로 잎의 병해 이미지를 획득함
- 잎에 발생하는 참외 노균병은 잎의 앞면에 잎맥을 경계로 하며 다각형의 반점으로 나타나고, 엷은 황색을 띄며, 병반 뒷면에 흰색 혹은 회색의 분생포자가 형성됨
- 잎의 잎맥을 경계로 부분적인 엷은 황색의 소반점이 나타나면 초기 피해로 간주함
-

- 잎에 발생한 소반점이 확대되며 황색의 다각형의 전형적인 병반이 잎 전반에 나타나면 중기 피해로 간주함
- 잎에 발생한 병반과 병반이 합쳐지며 건조해지면서 잎 전체가 안쪽으로 말리면서 고사하면 후기 피해로 간주함

11) 딸기흰가루병, Spharotheca humuli

- 주로 봄과 가을철에 발생하며 국내 주요 재배 품종인 설향은 내성으로 알려져 있으며 발생할 경우 봄에 발생량이 많고 잎과 잎자루 및 과육에 발생하여 피해를 일으킴.
- 순활물기생성 병원균으로 포장 내 잠재감염원으로부터 발생하므로 감수성인 킹스베리 품종을 사용하여 이병된 식물체를 이용한 인위접종을 통해 발병유도 예정
- 잎에 발생한 정도에 따라 시기별(발병초기, 발병중기, 발병말기)로 분류

12) 딸기잿빛곰팡이병, Botrytis cinerea

- 딸기의 잿빛곰팡이병은 대개 비교적 온도가 낮고, 높은 습도가 유지되는 환경 하에서 주로 발생함 시설을 이용한 반촉성 재배는 무가온으로 야간 또는 비가 내릴 때 밀폐가 되어 습도가 높아지며 특히 3-4월 사이에 비가 오면 토양수분의 상승으로 시설 안이 과습되어 발병에 이상적인 환경 조성되어 발생 함
- 처음에는 꽃잎에 침입하고 차차 수술, 암술, 꽃받침 및 수확기 과일에 발생하고 발생한 과일은 1-2일 사이에 문드러지고 회색가루의 곰팡이가 생기며, 화경, 화방경, 엽병 등은 감염되면 붉게 변하여 고사됨
- 꽃받침과 꽃을 중심으로 발생한 정도에 따라 시기별(발병초기, 발병중기, 발병말기)로 분류

13) 가지흰가루병, Golovinomyces cichoracearum

- 가지흰가루병은 전생육기에 걸쳐 잎, 줄기, 과실 등에 발생하며, 잎과 줄기에서 병징이 쉽게 확인이 가능하므로 잎과 줄기의 병해 이미지를 획득함
- 잎에 나타나는 가지흰가루병은 ❶원형의 흰 빛을 띠는 병반이 나타나고 ❷병반이 심해지면 조기 황화 낙엽현상이 나타남
- 원형의 흰 빛을 띠는 병반이 나타난 후 병이 심해지면 원형의 병반이 점점 잎이나 줄기 등에서 넓게 퍼지기 시작함, 원형의 흰 빛을 띠는 병반은 초기 피해로 간주함
- 원형의 흰 빛을 띠는 병반이 나타난 후 점점 병반이 넓게 퍼지게 되면 중기 피해로 간주함

- 잎 전체나 줄기 전체에 병반이 보이거나 조기 황화 낙엽현상이 보이면 말기 피해로 간주함

14) 가지잎곰팡이병, Mycovellosiella nattrassii

- 가지잎곰팡이병은 전생육기에 걸쳐 잎에 발생하므로 잎의 병해 이미지를 획득함
- 잎에 나타나는 가지잎곰팡이병은 ❶처음에는 잎 뒷면에 작은 반점이 형성되며 ❷병이 진전되면 가운데서부터 점점 회갈색으로 변하게 되어 원형의 그을음병반으로 확대되며 ❸확대된 후에는 연한 황갈색의 선명하지 않은 반점과 검은 균총이 나타남
- 잎 뒷면에 흰색의 곰팡이가 밀생하는 작은 반점이 형성되는데, 잎 뒷면의 작은 반점은 초기 피해로 간주함
- 그 후 병이 진전되며 원형의 그을음병반으로 확대되거나 엽맥부근엔 부정형으로 형성되면 중기 피해로 간주함
- 확대 된 후 병든 잎의 표면에 연한 황갈색의 선명하지 않은 반점과 원형의 균총이 보이면 말기 피해로 간주함

15) 상추균핵병, Sclerotinia minor

- 상추균핵병은 상추의 지제부에 병징이 먼저 나타나고 진전이 되면 그루 전체가 썩어서 쉽게 확인이 가능하므로 지제부와 그루 전체의 병해 이미지를 획득함
- 상추균핵병은 ❶지제부에 담갈색으로 물러썩는 형태와 ❷그루 전체가 썩어버리는 형태로 나타남
- 지제부에 담갈색의 썩음 증상 조금이라도 나타나면 초기 피해로 간주함
- 지제부의 썩음 증상이 심해져 상추잎까지 썩음 증상이 퍼지면 중기 피해로 간주함
- 그루 전체적으로 썩음 현상이 보이면 말기 피해로 간주함

16) 상추노균병, Bremia lactucae

- 상추노균병은 잎에 병징이 먼저 나타나고 쉽게 확인이 가능하므로 잎의 병해 이미지를 획득함
- 잎에 나타나는 상추노균병은 ❶잎에 작은 부정형 병반이 생기고 ❷병반이 심해져서 담갈색으로 변하고 잎 전체가 마름
- 잎 뒷면에 흰색 곰팡이가 생기고 잎 앞면에 부정형 병반이 나타나면 초기 피해로 간주함

- 부정형 병반이 퇴록반점으로 나타나고 병반이 합쳐져서 커지면 중기 피해로 간주함
- 잎 전체가 황록색 또는 황갈색으로 변하며 마르면 말기 피해로 간주함

17) 단호박흰가루병,Sphaerotheca fuliginea

- 주로 잎에 발생하며 잎자루와 줄기에도 발생한다. 잎에는 처음 흰색의 분생포자가 점점이 나타나고, 진전되면 잎 전체에 밀가루를 뿌려 놓은 것 같은 증상으로 변한다. 기온이 서늘해지면 병반상에 흑색소립점(자낭각)이 형성된다. 이 병으로 인하여 잎이 고사되는 일은 드무나 잎에 병든 그루는 노화되어 수확기간이 단축됨

18) 단호박점무늬병, Pseudomonas syringae pv. syringae

- 잎의 가장자리나 그 안쪽에 작고, 둥근 갈색병반이 생긴다. 병이 진전되면 병반은 동심원을 그리며 암갈색으로 확대되고, 나중에는 병반끼리 합쳐져서 찢어지고, 잎 전체가 낙엽 된다. 심한 경우 병반이 잎 반절이상을 차지하는 경우도 있다. 병반 표면에는 깨알 같은 흑색소립(병자각)이 생긴다.

19) 오이녹반모자이크바이러스, Cucumber green mottle mosaic virus

- 오이녹반모자이크바이러스병은 전생육기에 걸쳐 주로 잎에 병징이 뚜렷하게 나타나서 잎의 병해 이미지를 획득함
- 잎에 나타나는 오이녹반모자이크바이러스병은 ❶연녹색 반점, ❷모틀, ❸엽맥간 황화, ❹황화 현상이 나타남
- 연녹색 반점이 나타나면 초기 피해로 간주함
- 연녹색 반점 후, 엽맥간 황화 증상이 나타나면 중기 피해로 간주함
- 연녹색 반점 후, 모틀 증상이 나타나면 중기 피해로 간주함
- 엽맥간 황화 증상 후, 잎의 전반적인 황하 현상이 나타나면 말기 피해로 간주함

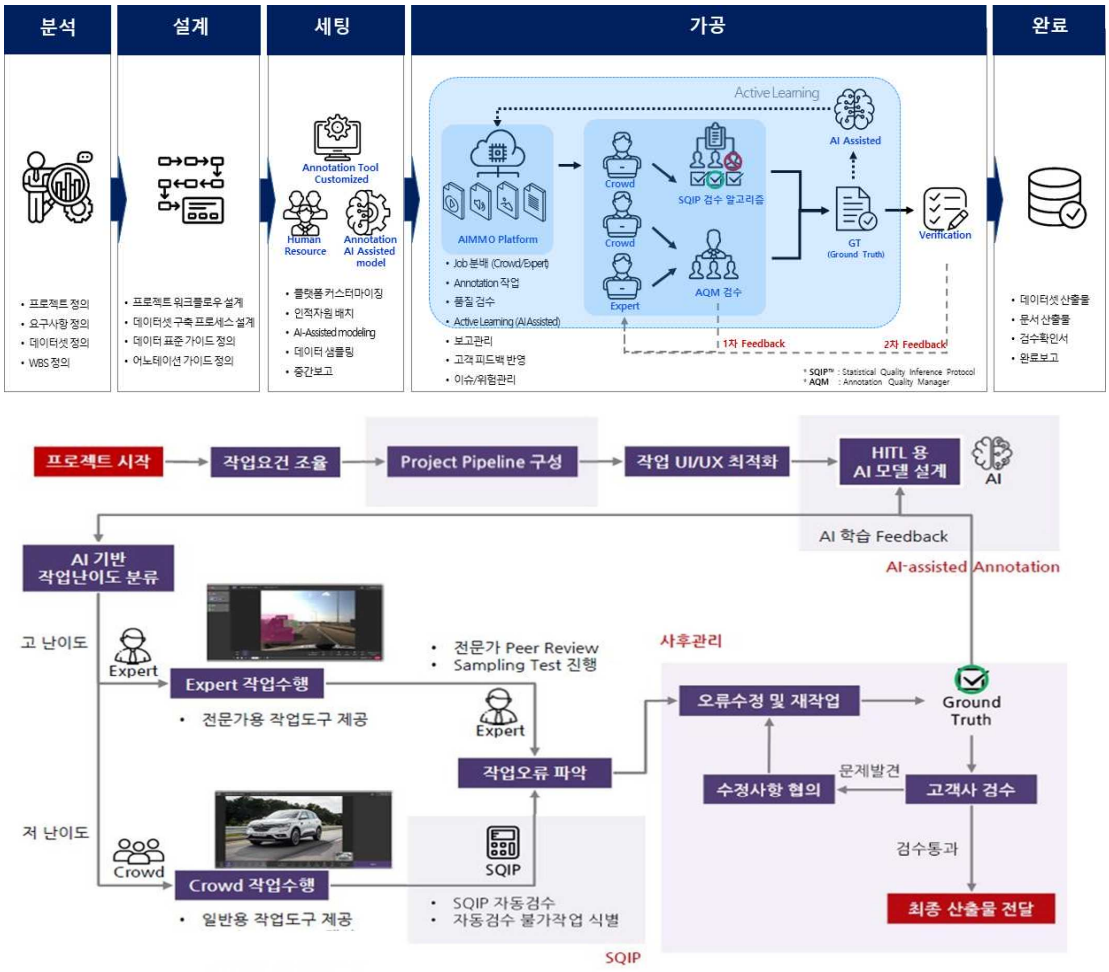
20) 애호박점무늬병, Pseudomonas syringae

- 잎의 가장자리나 그 안쪽에 부정형의 작거나 큰 갈색병반이 생긴다. 병이 진전되면 병반은 부정형의 암갈색에서 검은색으로 확대되고, 나중에는 병반끼리 합쳐져서 찢어지고, 잎 전체가 마른다. 심한 경우 병반이 잎 반절이상을 차지하는 경우도 있다.

3 어노테이션/라벨링

3.1 어노테이션 / 라벨링 절차

1) 라벨링 절차를 '분석', '설계', '세팅', '가공', '완료' 단계로 세분화 하여 추진



[그림 Ⅲ-135] 학습용 데이터 가공 프로세스

2) (분석) 상품이미지 데이터 어노테이션 프로젝트에 대한 정의 및 응용서비스 개발자의 요구사항 분석과 데이터셋의 정의를 수행하고 이에 따른 전체 가공작업의 WBS를 수립

3) (설계) 프로젝트 워크플로우를 하위 작업 단계(Stage)로 세분화하여 설정

- 본 작업의 특성을 고려한 최적의 가공 방식, 툴, 인력 구성, 목표 일정 등을 검토하고 데이터셋 어노테이션을 위한 최적의 프로세스 설계
- 데이터셋 어노테이션 가이드 정의

4) (세팅) 관리자, 검수자, 작업자 등 프로젝트 수행 구조 설정

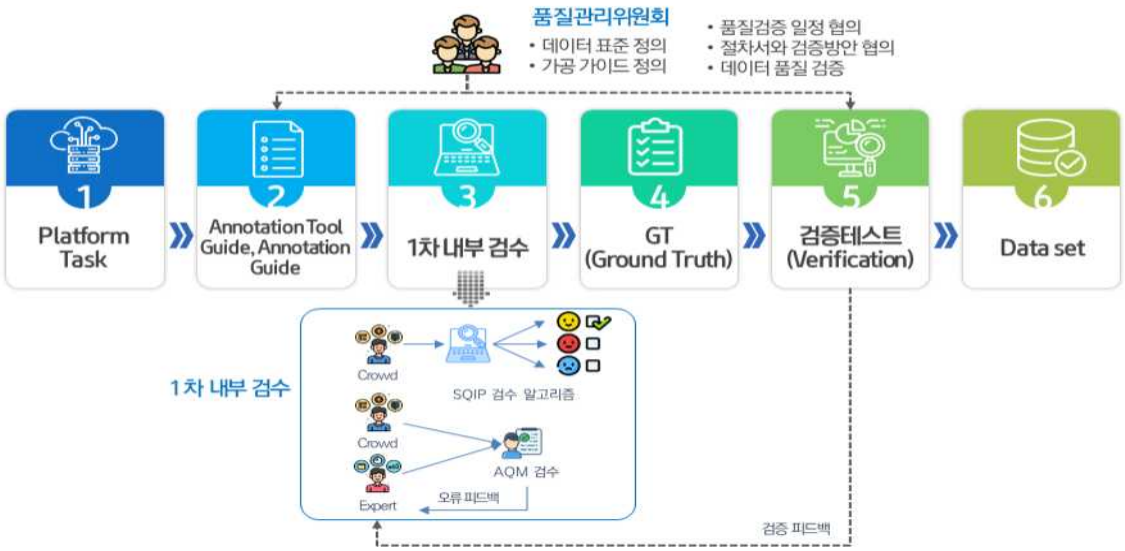
- 프로젝트에 적합한 어노테이션 플랫폼 커스터마이징 수행
- 작업자의 능력(정확도, 시간 등)을 토대로 작업자 선정 및 Task 할당 프로젝트 조직 설정 및 이메일을 통한 작업자 초대
- 프로젝트 어노테이션 유형, 분류 기준을 설정하고, 샘플 데이터를 통한 작업 시뮬레이션으로 자주 발생하는 오류 유형, 주의 사항 등을 사전에 파악하여 어노테이션 가이드라인에 반영 (어노테이션 가이드라인은 즉시 볼 수 있도록 화면에서 One-Click로 제공)
- 프로젝트 작업 이미지 포팅

5) (가공) 작업자들이 정확하고, 효율적으로 가공을 수행할 수 있도록 하는 효율적인 시스템 제공, 작업 공정 관리, 비효율에 대한 피드백 및 개선 등을 포함

- 작업자에 할당 된 Task를 실시간으로 확인(작업 전, 작업 중, 검수 중, 검수 완료) 및 진행
- 작업 환경내 커뮤니케이션 기능으로 변경 이슈 반영
- 주기적인 검수를 통해 빈번한 오류를 유형화하며 이를 가공 시스템 기능 개선(개발팀), 작업자 재교육(필요시), 오류 수정을 위한 재작업 등을 수행

6) (검수) 작업자들의 어노테이션 결과에 대한 전수 검수를 통한 데이터 품질 99% 이상 제공

- 데이터의 정확도 및 유효성 관점에서의 검증 계획서를 수립하고, 각 항목에 대한 목표치 설정 및 검증을 수행
- 각 항목에 대한 가공 품질의 목표치를 달성하기 위한 품질 제고 활동 수행



[그림 Ⅲ-136] 품질관리 활동

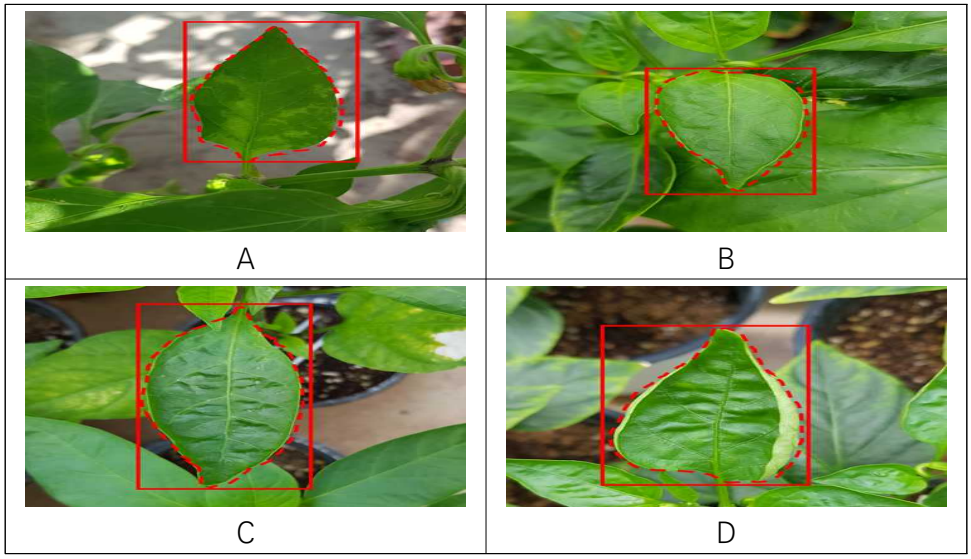
3.2 어노테이션 / 라벨링 기준

1) 데이터 작성(Annotation) 항목

- 원천 데이터 이름, 원천 데이터 해상도, 원천 데이터 증강(augmentation) 및 변형(색상, 기울기, 배경, 노이즈 등) 및 속성, 촬영(단일, 복수) 여부, 분류 코드, 상품 분류 값

2) 고추마일드모틀바이러스병, Pepper mild mottle virus

- 고추바이러스모틀바이러스병징의 경우 괴사 내지는 반점 등이 나타나지 않고 잎전반에 걸쳐 변형된 형태로 나타나므로 획득한 이미지에서 병징이 나타난 잎을 사각형 박스안에 위치 시키고 그림에서와 같이 잎 이미지의 가장자리를 따라 점선으로 윤곽을 표시함
- 그림 C에 나타난 것처럼 고추바이러스모틀바이러스병징이 나타난 잎이 다른 잎에 의해 겹쳐서 가린 경우에 가린 잎은 무시하고 병징 잎의 윤곽을 가상하여 표시함

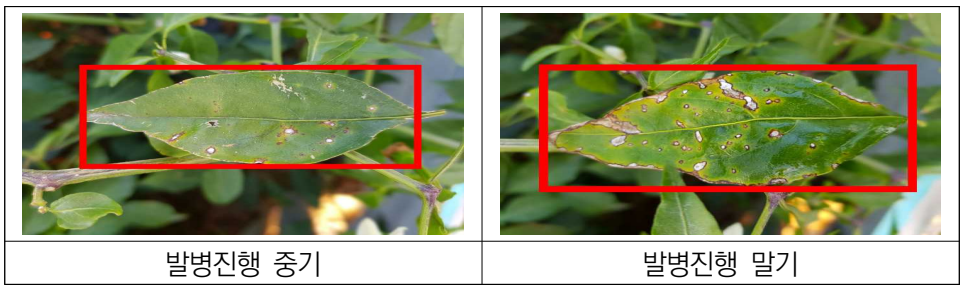


[그림 Ⅲ-137] 고추마일드모틀바이러스병 데이터 바운딩박스 및 윤곽

※ 획득한 이미지 내 병징이 다수 발생 또는 발생한 객체가 다수인 경우 하나의 박스만 생성함
※ 획득한 이미지에 박싱 및 라이닝의 1차 가공은 크라우딩소싱 요원들이 실시함
※ 크라우딩소싱 요원들에 의해 1차 가공된 이미지의 박싱 및 라이닝에 대한 QC 작업은 전문가에 의해 교육을 받은 QC 요원들이 전수 실시함
※ 획득한 발병 이미지에서 병의 발병정도는 전문가에 의해 교육 받은 QC 요원이 전문가의 도움을 받아가며 전수 확인함

3) 고추점무늬병, Xanthomonas campestris

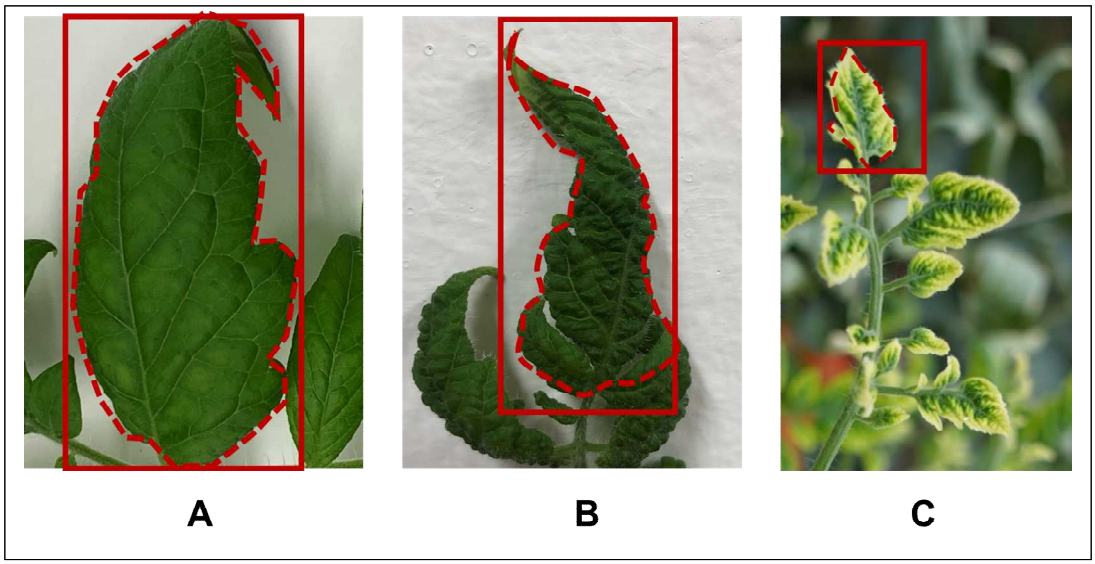
- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱작업을 통해 어노테이션을 실시함
- 고추 점무늬병의 획득한 이미지에서 병징부분을 박스처리하여 AI 인식부분임을 표기함
- 고추 점무늬병에 걸린 잎들이 다른 잎에 의해 겹쳐서 가린 경우에 가린 잎은 무시하고 병징 잎의 윤곽을 가상하여 표시함



[그림 Ⅲ-138] 고추점무늬병 데이터 바운딩박스

4) 토마토황화잎말이바이러스병, Tomato yellow leaf curl virus

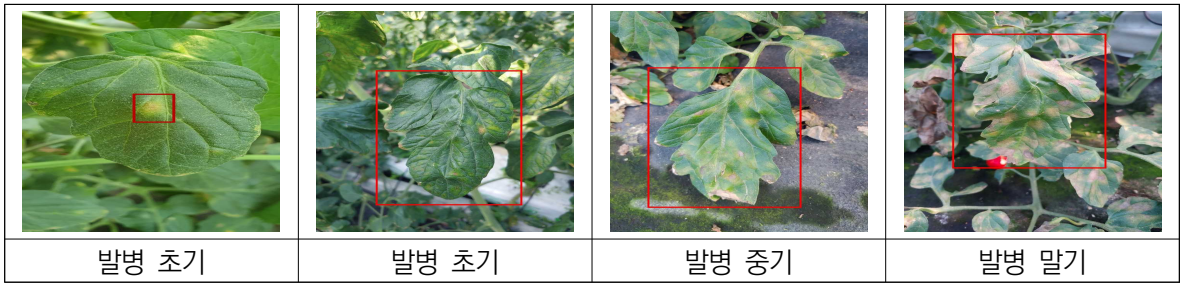
- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 토마토황화잎말이바이러스병징의 경우 괴사 내지는 반점 등이 나타나지 않고 잎전반에 걸쳐 변형된 형태로 나타나므로 획득한 이미지에서 병징이 나타난 잎을 사각형 박스안에 위치 시키고 그림에서와 같이 잎 이미지의 가장자리를 따라 점선으로 윤곽을 표시함
- 그림 C에 나타난 것처럼 토마토황화잎말리바이러스병징이 나타난 잎이 다른 잎에 의해 겹쳐서 가린 경우에 가린 잎은 무시하고 병징 잎의 윤곽을 가상하여 표시함



[그림 Ⅲ-139] 토마토황화잎말이바이러스병 바운딩박스 및 윤곽

5) 토마토잎곰팡이병, Fulvia fulva

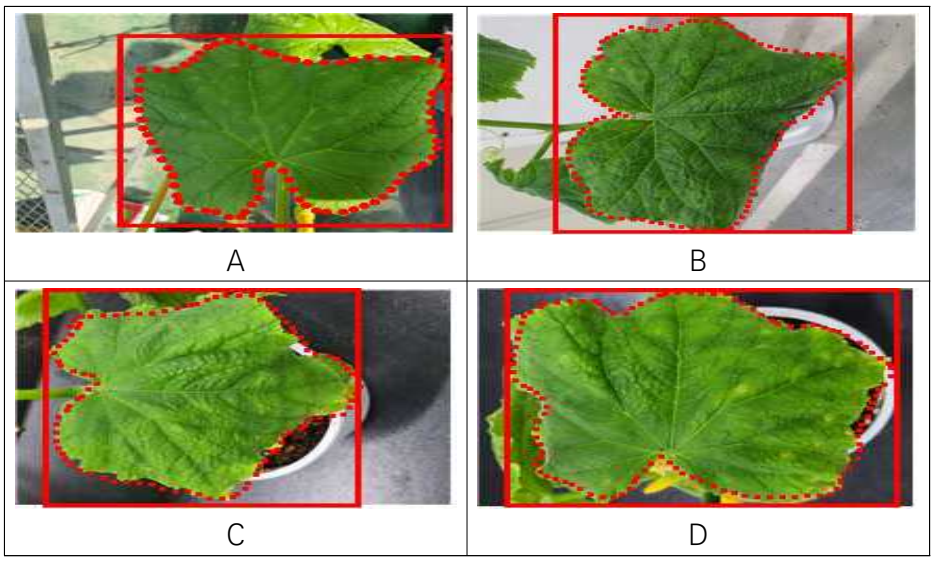
- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 토마토잎곰팡이병의 획득한 이미지에서 병징부분을 박스처리하여 AI 인식부분임을 표기함
- 이병된 잎의 앞면과 뒷면의 특징이 서로 달라 육안판별의 기준으로 삼긴하나, 본 과제에서는 앞면에 나타난 담황색의 병반을 인식 기준으로 선정함
- 토마토잎곰팡이병에 걸린 잎들이 다른 잎에 의해 겹쳐서 가린 경우에 가린 잎은 무시하고 병징 잎의 윤곽을 가상하여 표시함



[그림 Ⅲ-140] 토마토잎곰팡이병 데이터 바운딩박스

6) 오이녹반모자이크바이러스, Cucumber mosaic virus

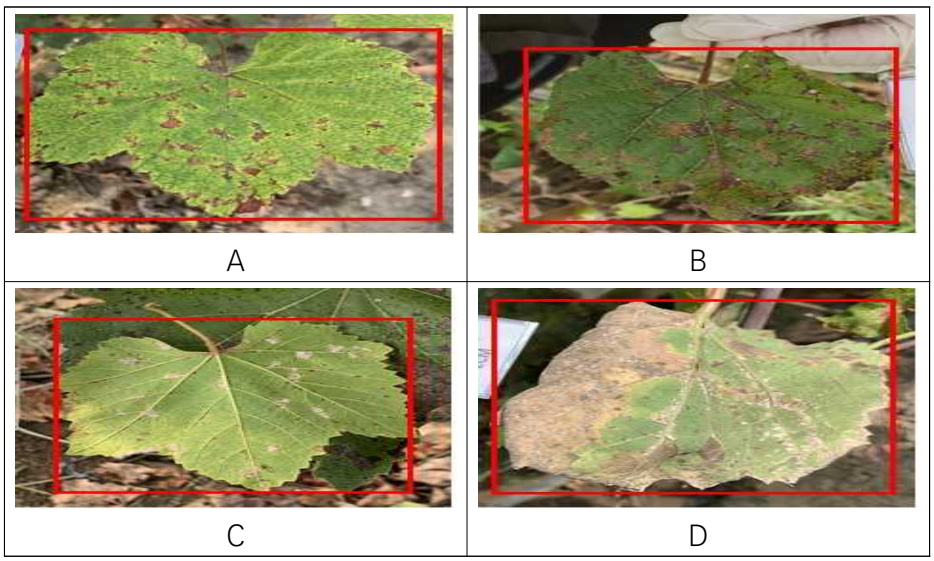
- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 오이녹반모자이크바이러스병징의 경우 연녹색 반점 및 황화 증상이 나타나므로 획득한 이미지에서 병징이 나타난 잎을 사각형 박스안에 위치시키고 그림에서와 같이 잎 이미지의 가장자리를 따라 점선으로 윤곽을 표시함
- 아래 그림의 D에서 나타난 것처럼 오이녹반모자이크바이러스병이 나타난 잎이 다른 잎에 의해 겹쳐서 가린 경우에 가린 잎은 무시하고 병징 잎의 윤곽을 가상하여 표시함



[그림 Ⅲ-141] 오이녹반모자이크바이러스 데이터 바운딩박스 및 윤곽

7) 포도노균병, Plasmopara viticola

- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 포도노균병은 잎전반에 걸쳐 병반이 나타나므로 획득한 이미지에서 병징이 나타난 부분을 사각형 박스안에 위치 시킴

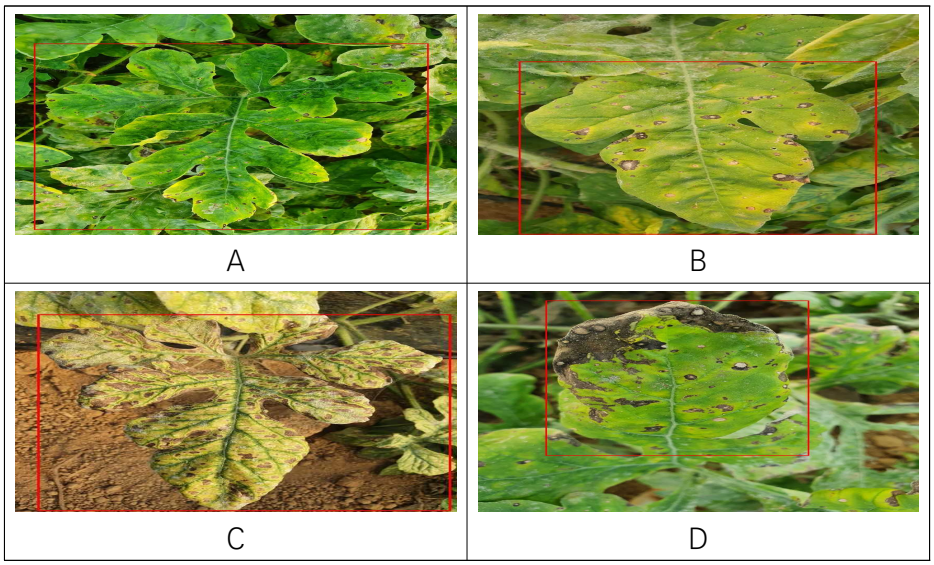


[그림 Ⅲ-142] 포도노균병 데이터 바운딩박스

8) 수박탄저병, Colletotrichum orbiculare

- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 수박 탄저병 병징의 경우 잎 전반에 걸쳐 불규칙한 반점으로 나타나므로 획득한 이미지에서 병징을 A 그림에서와 같이 병징이 나타나는 잎을 이미지 중앙에 위치시키고, 주변 잎의 간섭을 배제하기 위해 잎 이미지의 가장자리를 따라 점선으로 윤곽을 표시 및 탄저병 병반를 사각박스로 표시함
- 수박 잎 일부를 찍을 경우, 병징을 이미지 중앙에 위치시켜 B그림에서와 같이 잎 이미지의 가장자리를 따라 점선으로 윤곽을 표시 및 탄저병 반점 부위를 사각박스로 표시함
- 그림 C에 나타난 것처럼 병의 진전에 따라 수박 탄저병 병징이 겹쳐 병반이 확장되므로 잎 이미지의 가장자리를 따라 윤곽선으로 표시하고, 회흑색으로 변한 병반을 사각박스로 표시함

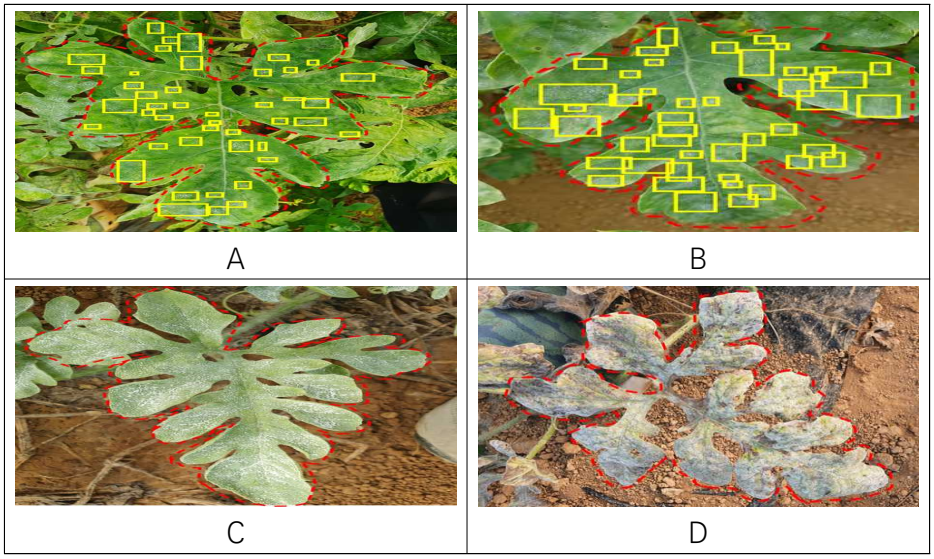
- 수박 탄저병 병징이 겹쳐 병반이 확장되어 있는 잎의 일부에서 찍을 경우, 수박 잎의 가장자리를 점선으로 윤곽을 표시하고, 그림 D에 나타난 것처럼 병반을 사각박스로 표시함
- 그림 B, C, D에 나타난 것처럼 수박탄저병 병징이 나타난 잎에 동시에 흰가루병 병징이 발생한 경우 흰가루병 병징은 무시하고 탄저병 병징의 윤곽을 가상하여 표시함



[그림 Ⅲ-143] 수박탄저병 데이터 바운딩박스

9) 수박흰가루병, Sphaerotheca fusca

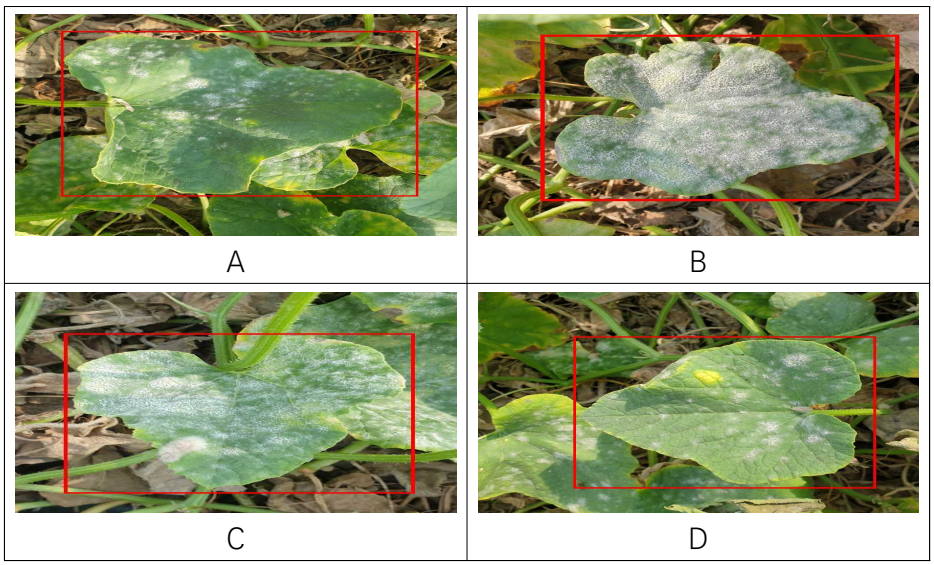
- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 수박 흰가루병 병징의 경우 병의 초중기에 잎전반에 걸쳐 불규칙한 원형으로 나타나므로 병징이 나타난 잎을 이미지 중앙에 위치시키고 주변부 이미지의 간섭을 배제하기 위해 그림에서와 같이 잎 이미지의 가장자리를 따라 점선으로 윤곽을 표시함
- 수박 잎 일부를 찍을 경우, 병징을 이미지 중앙에 위치시켜 B그림에서와 같이 잎 이미지의 가장자리를 따라 점선으로 윤곽을 표시 및 흰가루 병총 부위를 사각박스로 표시함
- 그림 C에 나타난 것처럼 수박 흰가루병 병징이 잎 전체를 백색 또는 회백색의 병반을 덮으면 이미지의 가장자리를 따라 윤곽선으로 표시함
- 그림 D에 나타난 것처럼 수박 흰가루병 병징이 나타난 잎에 동시에 수박 탄저병 병징이 발생한 경우 탄저병 병징은 무시하고 흰가루병 병징의 윤곽을 가상하여 표시함



[그림 Ⅲ-144] 수박흰가루병 데이터 바운딩박스 및 윤곽

10) 참외흰가루병, Sphaerotheca fusca

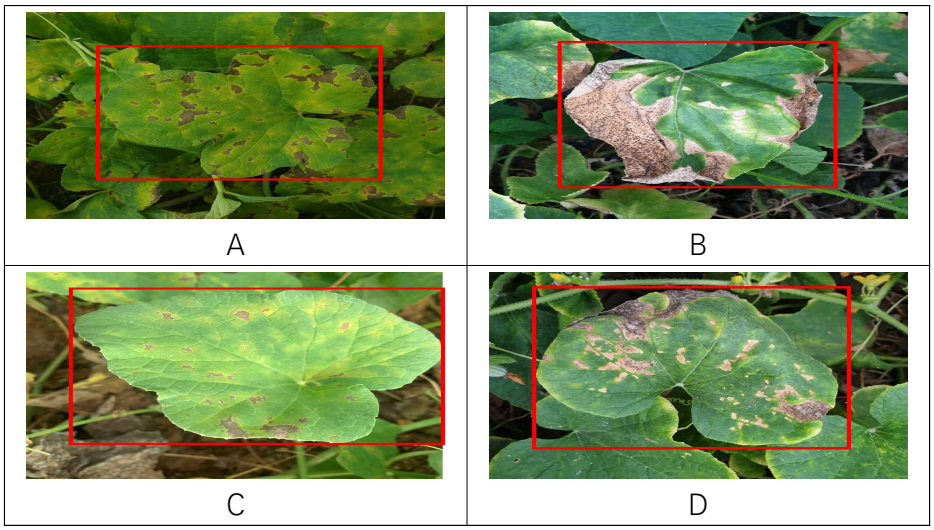
- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 참외 흰가루병의 병징의 경우 잎 전반에 걸쳐 흰색의 반점이 관찰됨으로 획득한 이미지에서 병징이 나타난 사각형 박스안에 위치시킴



[그림 Ⅲ-145] 참외흰가루병 데이터 바운딩박스

11) 참외노균병, Pseudoperonospora cubensis

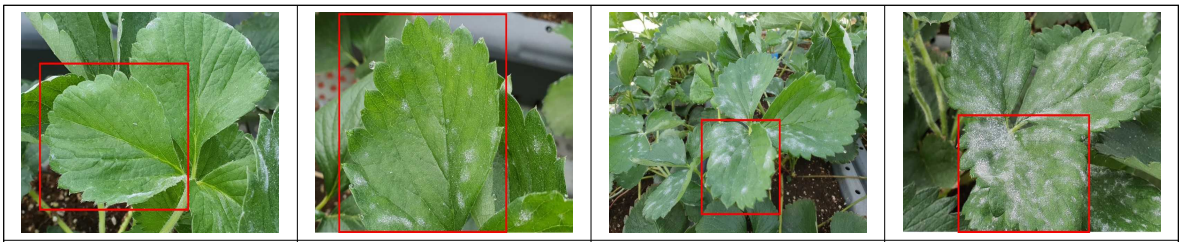
- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 참외 노균병의 경우 엷은 황색의 다각형의 반점이 잎 전반에 걸쳐 나타나므로 획득한 이미지에서 병징이 나타난 잎을 사각형 박스 안에 위치시킴



[그림 Ⅲ-146] 참외노균병 데이터 바운딩박스

12) 딸기흰가루병, Spharotheca humuli

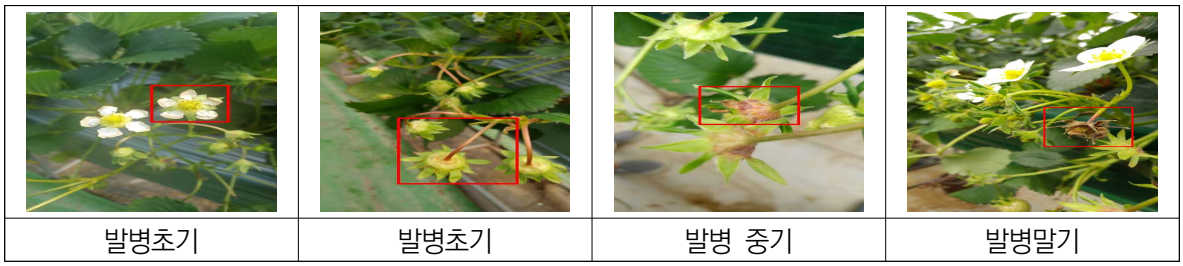
- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 딸기흰가루병의 획득한 이미지에서 병징부분을 박스처리하여 AI 인식부분임을 표기함
- 토마토잎곰팡이병에 걸린 잎들이 다른 잎에 의해 겹쳐서 가린 경우에 가린 잎은 무시하고 병징 잎의 윤곽을 가상하여 표시함



[그림 Ⅲ-147] 딸기흰가루병 데이터 바운딩박스

13) 딸기잿빛곰팡이병, Botrytis cinerea

- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 딸기잿빛곰팡이병의 획득한 이미지에서 병징부분을 박스처리하여 AI 인식부분임을 표기함
- 딸기잿빛곰팡이병에 걸린 꽃받침들이 다른 꽃받침에 의해 겹쳐서 가린 경우에 가린 꽃받침은 무시하고 병징 잎의 윤곽을 가상하여 표시함



발병초기	발병초기	발병 중기	발병말기

[그림 Ⅲ-148] 딸기잿빛곰팡이병 데이터 바운딩박스

14) 가지흰가루병, Golovinomyces cichoracearum

- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 가지흰가루병의 잎의 경우 병반이 잎의 일부 또는 전체에 나타나므로 획득한 이미지에서 병징이 나타난 잎을 사각형 박스 안에 위치시키고 그림과 같이 잎 전체 부분을 따라 점선으로 윤곽을 표시함
- 가지흰가루병의 줄기의 경우 병반이 줄기 일부 또는 전체에 나타나므로 획득한 이미지에서 병징이 나타난 줄기를 사각형 박스 안에 위치시키고 그림과 같이 줄기 전체 부분을 따라 점선으로 윤곽을 표시함

A	B

[그림 Ⅲ-149] 가지흰가루병 데이터 바운딩박스 및 윤곽

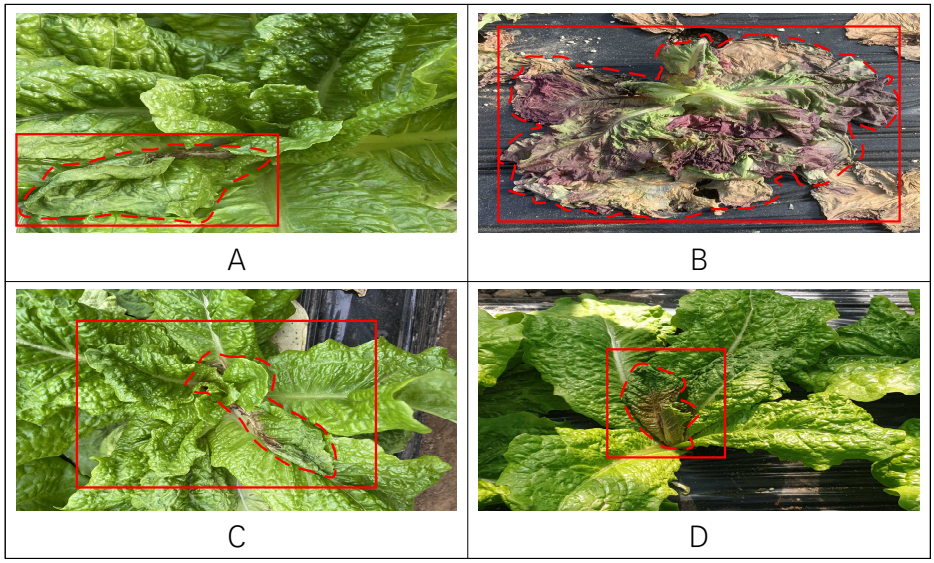
15) 가지잎곰팡이병, Mycovellosiella nattrassii

- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 가지잎곰팡이병 병징의 경우 잎의 넓은 부위에 나타나므로 획득한 이미지에서 병징이 나타난 잎을 사각형 박스안에 위치 시키고 그림에서와 같이 잎 이미지의 가장자리를 따라 점선으로 윤곽을 표시함

[그림 Ⅲ-150] 가지잎곰팡이병 데이터 바운딩박스 및 윤곽

16) 상추균핵병, Sclerotinia minor

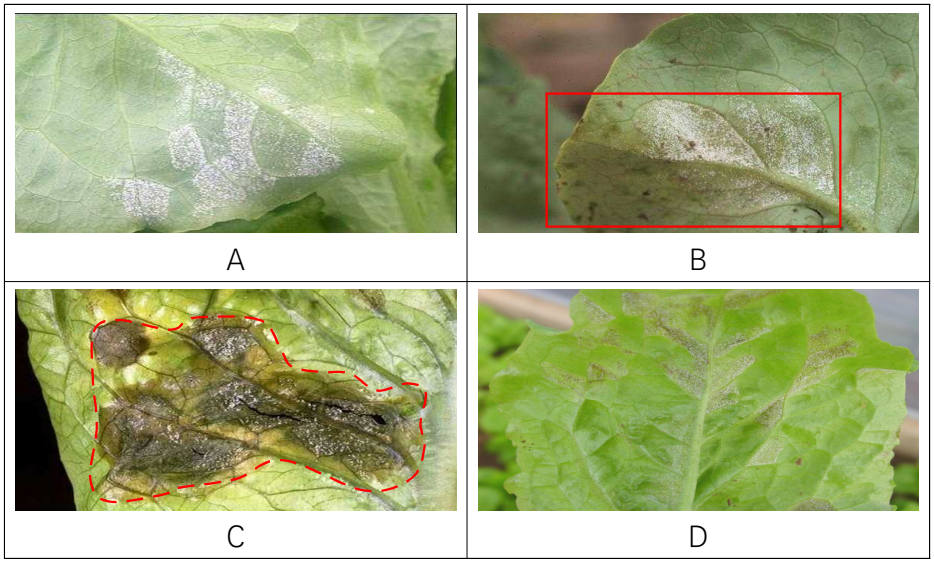
- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 상추균핵병징의 경우 상추 지제부와 잎에 썩음이 일어나므로 획득한 이미지에서 병징이 나타난 지제부 부분 또는 잎을 박스안에 위치 시키고 그림에서와 같이 병이 난 지제부 부분 또는 잎을 따라 점선으로 윤곽을 표시함
- 그루 전체에 썩음 현상이 있는 경우 획득한 이미지에서 병징이 나타난 그루 전체를 박스안에 위치 시키고 그림에서와 같이 그루의 가장자리를 따라 점선으로 윤곽을 표시함



[그림 Ⅲ-151] 상추균핵병 데이터 바운딩박스 및 윤곽

17) 상추노균병, Bremia lactucae

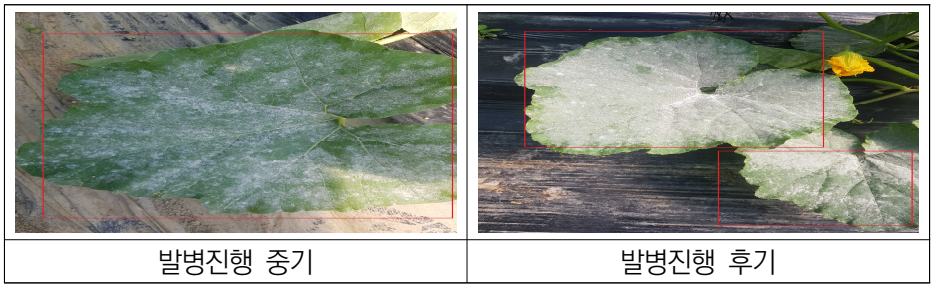
- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 상추노균병병징의 경우 부정형 병반이 잎의 일부에 나타나므로 획득한 이미지에서 병징이 나타난 잎을 사각형 박스안에 위치 시키고 그림에서와 같이 잎의 병반부분을 따라 점선으로 윤곽을 표시함
- 상추노균병병징이 심해져서 잎 전체적으로 마름현상이 일어나면 획득한 이미지에서 병징이 나타난 잎을 사각형 박스안에 위치 시키고 그림에서와 같이 잎의 가장자리를 따라 점선으로 윤곽을 표시함



[그림 Ⅲ-152] 상추노균병 데이터 바운딩박스 및 윤곽

18) 단호박흰가루병,Sphaerotheca fuliginea

- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 단호박 흰가루병의 획득한 이미지에서 병징부분을 박스처리하여 AI 인식부분임을 표기함
- 단호박 흰가루병에 걸린 잎들이 다른 잎에 의해 겹쳐서 가린 경우에 가린 잎은 무시하고 병징 잎의 윤곽을 가상하여 표시함

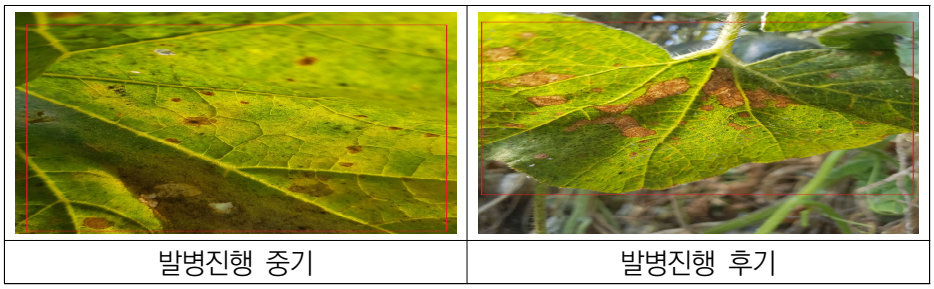


[그림 Ⅲ-153] 단호박흰가루병 데이터 바운딩박스

19) 단호박점무늬병, Pseudomonas syringae pv. syringae

- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 단호박 점무늬병의 획득한 이미지에서 병징부분을 박스처리하여 AI 인식부분임을 표기함

- 단호박 점무늬병에 걸린 잎들이 다른 잎에 의해 겹쳐서 가린 경우에 가린 잎은 무시하고 병징 잎의 윤곽을 가상하여 표시함

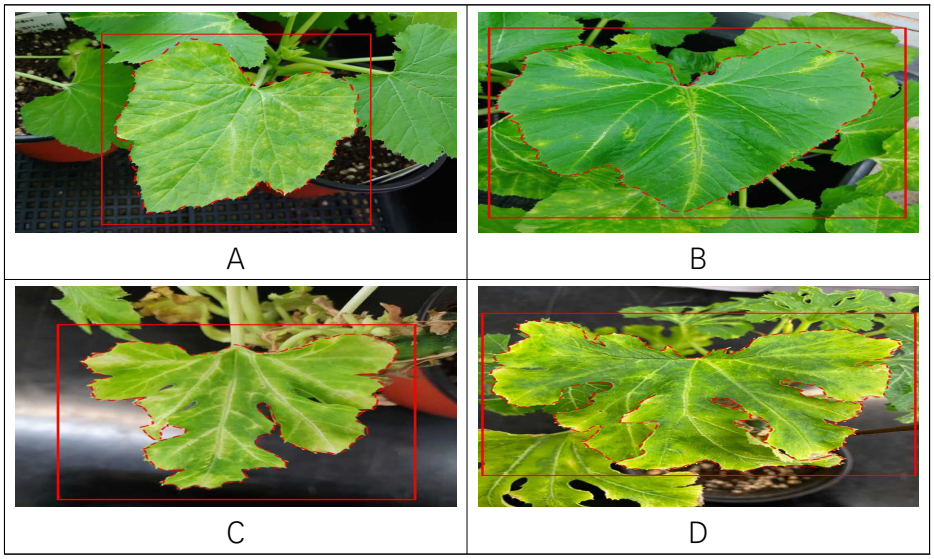


발병진행 중기	발병진행 후기

[그림 Ⅲ-154] 단호박점무늬병 데이터 바운딩박스

20) 오이녹반모자이크바이러스, Cucumber green mottle mosaic virus

- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 오이녹반모자이크바이러스병징의 경우 잎전반 걸쳐 연녹색 반점, 모틀, 엽맥간 황화, 황화로 나타나므로 획득한 이미지에서 병징이 나타난 잎을 사각형 박스안에 위치 시키고 그림에서와 같이 잎 이미지의 가장자리를 따라 점선으로 윤곽을 표시함
- 그림 B에 나타난 것처럼 오이녹반모자이크바이러스병징이 나타난 잎이 다른 잎에 의해 겹쳐서 가린 경우에 가린 잎은 무시하고 병징 잎의 윤곽을 가상하여 표시함



A	B
C	D

[그림 Ⅲ-155] 오이녹반모자이크바이러스 데이터 바운딩박스 및 윤곽

21) 애호박점무늬병, Pseudomonas syringae

- 획득한 병해 이미지를 병해 진전 정도에 따라 구분하고 병해 특징이 나타나도록 박싱과 라이닝작업을 통해 어노테이션을 실시함
- 애호박점무늬병의 획득한 이미지에서 병징부분을 박스처리하여 AI 인식부분임을 표기함
- 애호박점무늬병에 걸린 잎이 병이 크게 진전되어 잎의 과반이상이 검은색이 되거나 검은 점무늬가 불규칙하게 많이 나타난 병반은 병징이 나타난 부분을 큰박스 처리를 하여 표시함

	
발병진행 중기	발병진행 후기

[그림 Ⅲ-156] 애호박점무늬병 데이터 바운딩박스

3.3 어노테이션 / 라벨링 교육

- 크라우드 작업자 난이도별(초급/중급/고급) 온/오프라인 교육 실시
- 교육 및 작업 결과에 따라 프로젝트 수행 후 검수자, 내부 관리자로 채용 기회 제공

3.4 어노테이션 / 라벨링 도구 및 사용법

1) 어노테이션 전문기업이 개발한 저작도구 활용

2) (작업) 작업자 활용을 위한 편의 기능 탑재(작업이동, 화면 조절 등)

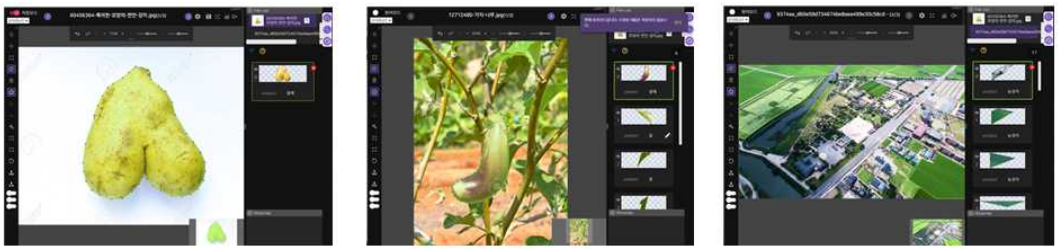
① (작업 이동) 작업後, 작업中, 작업前 파일을 간편을 간편하게 이동

② (화면 조절) 화면 상단에서 작업 화면 크기/밝기/색상, 실행 취소, 재실행 등 제공

③ (가공 지원) 어노테이션 작업 편의를 위한 간편하게 클릭만으로 이미지 이동, 바운딩박스, 라벨 ,선택 삭제, 초기화 기능 등을 제공

3) (결과물) 작업 결과물을 다양한 형식 제공할 수 있도록 데이터 포맷 컨버터 기능 지원

4) (저작도구 공개) 과제를 통해 진행된 저작도구는 소스와 기술 매뉴얼(데이터셋 형태, 규모, 특성 등)을 공개하여 외부에서 활용이 가능



[그림 Ⅲ-157] 데이터 저작도구 활용 화면

4 데이터 검수

4.1 검수 절차

- 검수를 위해 어노테이션 공정별 6단계 절차로 검수 진행(2인 이상 수작업 검증 실시)
- 공정별 품질관리 목표를 설정하고 진단·개선을 통해 고품질 AI 학습용 데이터 제작

〈표 Ⅲ-118〉 데이터 품질관리 프로세스

구 분	프로세스	설명
데이터 분석	대상 식별	• 고객사의 품질관리 요구사항을 확인 • 품질관리를 수행할 대상을 구체화 및 문서화
▼		▼
데이터 설계	규칙 정의	• 품질관리 대상에 대한 프로파일링을 시행하고 • 품질 측정 및 통제를 위한 지표를 설정 • 설정된 품질규칙은 데이터 가공 업무규칙에 반영
▼		▼
데이터 가공	측정	• 데이터 가공 결과물 중 품질관리 대상에 대한 품질 측정
	분석	• 품질 측정 결과를 품질지표와 비교하여 시사점 도출 • 개선이 필요한 부분에 대한 원인 및 개선방법 분석
	개선	• 오류의 영향도 및 시급성을 고려하여 개선 시행
	통제	• 품질측정-분석-개선이 선순환 구조를 이룰 수 있도록 지속적인 모니터링 수행

4.2 검수 기준

- 데이터 검수 기준

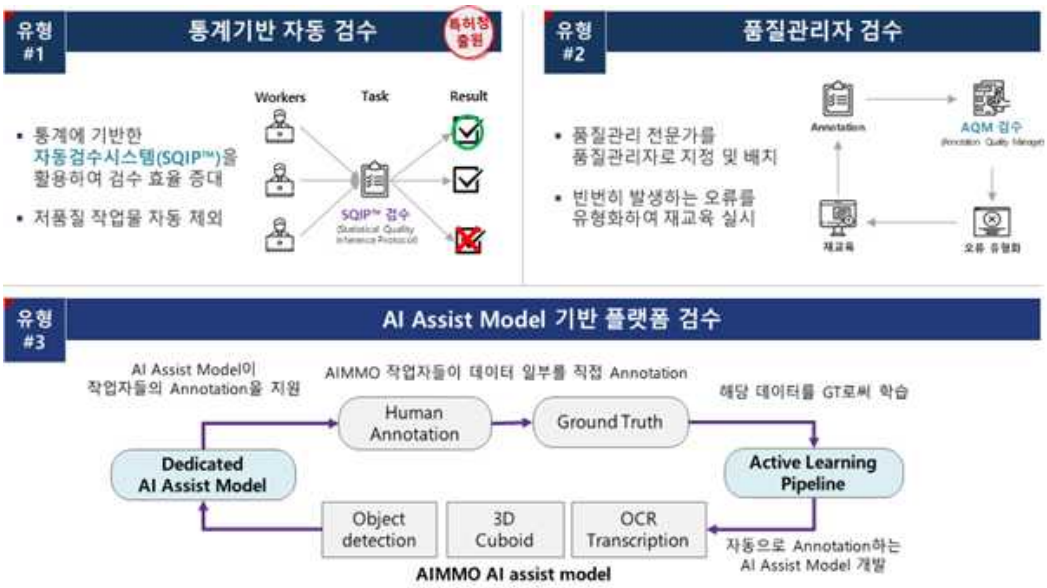
〈표 Ⅲ-119〉 데이터 검수 기준

품질특성	항목명	측정 지표	정량 목표
다양성	정상 및 각 질병 이미지	이미지 수	정상작물별 10,000 장, 질병별 1,000 장
	질병 단계별 이미지	각 질병 단계별 이미지 비율	단계별 10 % 이상
정확성	라벨링 구문적 정확도	정확도	99 % 이상
	라벨링 정확도	정밀도	90 % 이상
유효성	질병 인식 정확도	F1-score	0.7 이상

- 검수 조직 및 검수 도구

〈표 Ⅲ-120〉 데이터 유형별 검수 방법

유형	구성	검수 방법
통계기반 자동 검수	크라우드 소싱 인력	통계 기반의 자동 검수 시스템(SQIP)을 활용 자동 검수를 통한 저품질 작업물 제외하여 효율 증대
품질관리자 검수	품질 전문가 및 재택 전문 작업자	프로세스 및 데이터 별 전수 검사 빈번한 오류 발생 시 유형화 하여 프로세스 개선



[그림 Ⅲ-158] 데이터 유형별 품질검수

- 기타 품질관리 활동
  - (외부 품질관리) 산학연관 데이터 품질관련 전문가로 구성된 '품질관리위원회'를 구성하여 데이터 어노테이션 전 과정에 대한 품질공정으로 데이터 순도 극대화
  - (TTA 품질검증) 학습데이터 품질검증에 필요한 자료 및 환경(도구) 제공 및 적극 지원

〈표 Ⅲ-121〉 TTA 품질검증 체계별 지원사항

구분	구축공정	정확도	유효성
검증대상	공정 전주기	데이터 및 저장소	학습모델
검증방법	• 문서 검토 • 수행기업 인터뷰 • 현장 점검, 자료 확인	• 전수 또는 샘플링 검사 • 자동화 검수 도구 • 검증 데이터 분석	• 학습조건 설정 및 수행 (데이터 구분, 반복 횟수 등)
지원사항	• 작업공정도 제공 • 인터뷰/현장점검 지원	• 품질검증 데이터 제공 • 저작도구 활용 지원	• AI 학습결과 검증 지원
확인	품질검증 결과서(구축공정, 정확도, 유효성)		

5 데이터 활용 방안

5.1 학습 모델

1) (개발 목표) 작물의 질병해충 진단을 위한 AI 학습용 빅데이터 기반 인공지능 데이터 활용 모델 및 응용서비스 개발

2) (개발내용) 작물의 이미지를 활용하여 질병해충 최종진단을 위한 모바일/Web 응용서비스 제공을 위해 관련 최신기술을 적용 AI 학습용 모델을 개발하여 적용
   - Classification AI 모델 개발
   - 학습데이터 활용을 위한 인공지능모델 제공 및 개발 가이드라인 제공
   - 외부사용자 활용/지원을 위한 서비스모델(예: 개발지원 웹) 개발
   - 제품 개발 방안(SW): 플랜트 닥터 (Plant Doctor)
   - 작물 질병해충(충해포함) 학습을 위한 객체별(정상작물, 질병, 해충) 분류 데이터 DB구축

3) 샘플 알고리즘 요약

〈표 Ⅲ-122〉 샘플 알고리즘 요약

Task 분류	Task 상세	샘플 알고리즘	프레임워크
Classification	• 작물의 구분 • 해당 작물의 질병 정보 구분	ResNet50	PyTorch (.pt, .pth)

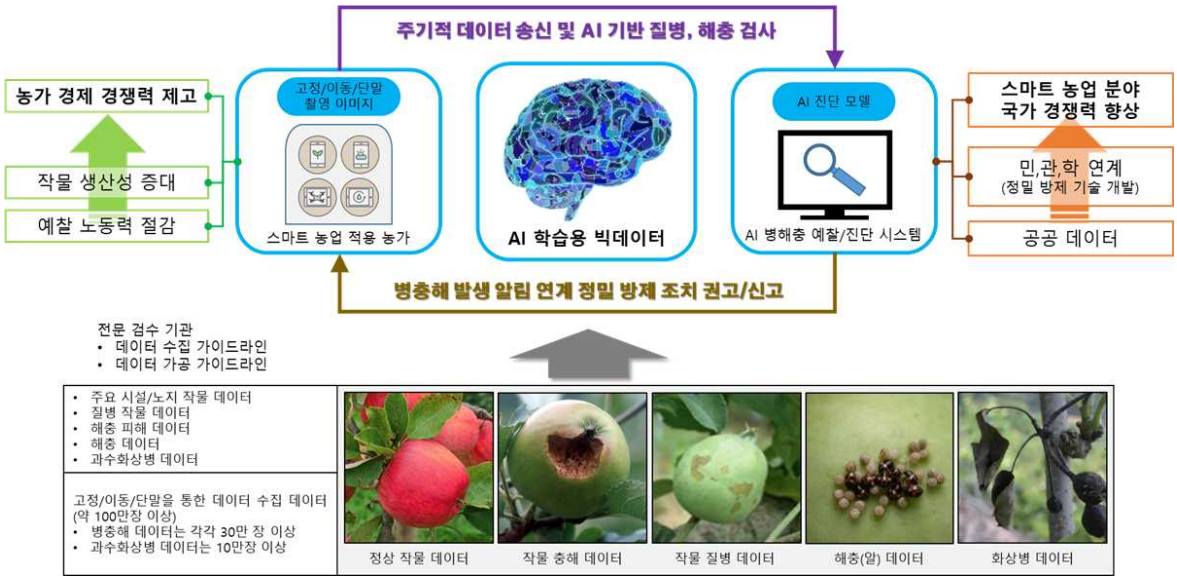
4) Classification

- Image classification이란 주어진 이미지를 학습한 class들 중 하나로 분류하는 task 임. 이는 이미지에 하나의 물체가 있을 때, 이 물체가 학습한 class중 어떠한 class에 속하는지를 분류하기 위함임. 그러나 하나의 이미지에 여러 개 class의 object가 동시에 존재할 때에는 사용하기 부적절함
- 데이터베이스 구축 시 각 이미지에 해당하는 작물, 부위, 질병의 정보를 포함함으로써 작물별, 부위별, 질병별 class를 구분하는 알고리즘 개발이 가능하도록 함. 이를 위해 이미지 전체에 대한 class 구분을 할 수도 있으나, 더 나아가 주목 객체의 bounding box를 제공함으로써 특정 주목 부위에 대한 class를 구분하는 알고리즘 개발 또한 가능함
- 샘플 모델에 대해서는 training set과 testing set을 명확히 구분하여 학습을 진행하고 해당 모델에 대한 accuracy를 제공함

〈표 Ⅲ-123〉 수박탄저병 classification 예시

이미지	질병 정보
	• 수박탄저병 : 90% • 수박흰가루병 : 10% • 정상 : 0%

5.2 서비스 활용 시나리오



[그림 Ⅲ-159] 작물의 질병해충 AI 모델 고도화 서비스

[그림 Ⅲ-160] 서비스 활용 시나리오 예시

[그림 Ⅲ-161] 시설작물 질병진단 서비스

- 시설에서 작물을 재배중인 농민은 모바일 앱을 통해 주기적으로 혹은 수시로 질병이 의심되는 부위를 촬영하여 이미지를 업로드함. 우선적으로는 농가가 업로드 한 이미지를 분석하여 의심되는 질병을 진단하여, 모바일앱을 통해 질병 진단결과 및 상황 조치에 필요한 농약과 처리 방법을 해당 농가에게 전달 함
- 모바일 앱을 통해 농가에 전달되는 조치사항은 PLS기준에 맞도록 구성 되어 있으며, 만약 구입해야 할 농약의 재고가 없어 구매가 필요할 경우 해당 질병의 치료에 필요한 약품을 자동 주문할 수 있는 구매 서비스를 함께 제공함 (응용서비스의 구현에 있어 본 구매기능은 가상의 상황으로써 구현함
- 질병상황 조치 이후 모바일 앱은 해당 질병이 잘 치료가 되었는지 확인이 될 때까지 알림 기능을 통해 해당 작물의 질병 부위에 대한 촬영 및 업로드를 농가에 요구함으로써 해당 질병의 치료가 최종 완료되었음을 확인함

제13장 동의보감 약초 이미지 AI 데이터

1 데이터 정보 요약

1.1 가이드 분류

대분류	이미지	중분류	비전	소분류	JPG

1.2 데이터 정보

데이터 이름	동의보감 약초 이미지 AI 데이터
활용 분야	농업, 교육, 생활안전, 국민건강, 연구, 표준화, 콘텐츠
데이터 요약	해마다 독초를 약초로 오인, 음용하여 발생하는 독초에 대한 중독 및 사망 사고를 감소시키고, 전문가 자문 없이 남용하여 몸을 해치지 않도록 약초 사용 방법 제공을 위해 동의보감 수록 457종 중 선정 약초 50종의 이미지 데이터와 약초의 올바른 사용 방법 제공에 필요한 약초 정보 등이 포함된 동의보감 약초 판별 AI 학습용 데이터

1.3 데이터 구축 개요

〈표 Ⅲ-124〉 단계별 데이터 구축 개요

단계	과정	내용
1단계	데이터 수집	• 선정된 약초 50종, 비교식물 78종에 대한 이미지 데이터 촬영은 최종 서비스 사용자의 시각을 고려하여 다양한 각도로 촬영 • 촬영 시 약초 전문가 동행 • 촬영 매뉴얼에 따라 선정 약초와 비교식물을 촬영하되 현장 상황 및 계절적인 요인으로 촬영 매뉴얼 적용이 어려운 경우 탄력적으로 대상 식물 촬영 ※ 대상 뿌리채취가 불가능하거나, 꽃이나 열매가 없는 등 현장 상황과 계절적인 요인으로 촬영이 어려운 경우를 말함

단계	과정	내용
2단계	데이터 정제 및 검증	• 수집된 데이터 분류 시 약초와 독초 이미지가 섞이거나 다른 폴더에 잘못 분류되지 않도록 담당자 사전 교육 시행 • 수집된 데이터는 인공지능 학습용 데이터 품질 요구에 충족될 수 있도록 기술적 검수, 전문가 검증 과정을 거침, • 특히 전문 검증 과정에는 한의학자 또는 식물 분류학자를 활용하여 약초와 독초가 섞이지 않도록 명확한 검증 과정 수행
3단계	데이터 가공 및 검수	• 데이터 품질 기준에 부합된 정제 데이터는 Annotation, Labeling 등의 과정을 통해 데이터 가공 • 데이터 가공은 약초 판별 인공지능 학습에 적합한 형태로 가공 • 가공된 데이터는 라벨링, 어노테이션 등에 대한 오류 검수
4단계	가공 데이터 활용	• 가공 데이터는 검수 과정을 통해 약초 판별 인공지능 학습모델 훈련에 활용되며 인공지능 연구 및 콘텐츠 제작 등 다양한 연구·개발용으로 공개됨

1.4 구축 목적

- 독초와 약초 구별에 도움이 되는 인공지능 학습모델 개발과 약초의 올바른 사용 방법 제공을 위한 데이터를 구축하여 일반 국민 누구나 쉽게 활용이 가능한 인공지능 응용 서비스 개발과 다양한 인공지능 연구에 활용할 수 있도록 하기 위함

1.5 활용 분야

- 약초 판별 알고리즘 및 인공지능 학습모델 개발
- 독초를 약초로 오인하여 발생 되는 중독 및 사망 사고 예방 안 제시

1.6 유의 사항

- 촬영 이미지 저작권 사용 허락 동의 여부 확인, 저작권 사용 허락 동의에 대한 법적 고지
- 촬영 이미지 데이터 저작권 활용 동의 과정 필수 수행
- 지적 저작권이 있는 식물의 경우 저작권 활용 동의 문서 작성 필수
- 농가 등 재배 약초의 경우 약초 구입 시 저작권 동의 및 개인정보 활용, 개인정보처리 방침 등 관련 동의 문서 작성
- 약초 구입 시 비용 지급은 근거를 남길 수 있도록 통장 입금 원칙
- 유사 약초와 섞이지 않도록 데이터 분류 및 허브넷 플랫폼 등록 시 특히 주의
- 등록 데이터는 전문가(한의학, 식물 분류학)의 검증 과정 필수 수행

2 데이터 획득 및 정제

2.1 원시 데이터 선정

- 동의보감 수록 약초 중에서 한약 처방 빈도가 높은 품목
  ※ 한약 처방 시 최소 50회 이상 사용(120종)
- 농림부 통계《약용작물 생산 실적에 집계된 국내 생산 약용작물
  ※ 농림축산식품부 2019년 약용식물 생산량 높은 작물(상위 62종)
- 국내 한약 산업 보호를 위해 수입 허가를 받아야 하는 '수급 조절 한약재' 품목
  ※ 11종(구기자, 당귀, 맥문동, 산수유, 오미자, 일당귀, 작약, 지황, 천궁, 천마, 황기)
- 국내에서 재배되고 있는 약용작물 중 계절적 요인이 적은 약초
  ※ 사업수행 시기를 고려하여 발아기, 개화기 이외에 계절적 제약 요인이 적은 약초 선정

[그림 Ⅲ-162] 원시데이터 선정 기준

- 선정약초의 약용 부위 : 약초는 약용 부위가 저마다 달라 촬영 시 약용하는 부위(6부위)를 중심으로 나눠 집중 촬영

〈표 Ⅲ-125〉 약초 약용 6부위

약초(식물)의 모양	약용 부위(명칭)	설 명
	전초(지상부)	약초 전체 또는 뿌리를 제외한 나머지를 지상부라 함
	잎	초본 또는 목본의 잎을 약용
	줄기(수피류)	초본 또는 목본의 줄기와 수피를 약용
	뿌리(뿌리 줄기)	뿌리 및 뿌리줄기를 약용
	꽃	꽃을 약용
	열매 및 종자류	열매와 종자 부분을 약용

2.2 규제관련 사항

- 관련 내용 없음

※ 본 데이터 가이드라인 상에는 규제 관련 내용이 없지만, 데이터 구축 시 규제 관련 내용 확인 및 검토 필요

2.3 획득 및 정제 절차

〈표 Ⅲ-126〉 약초 데이터 획득 및 정제 절차

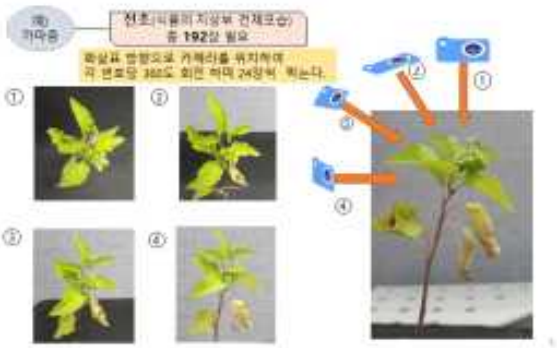
구분	내용
데이터 획득	① 대상 약초 50종 구분 및 촬영 방법 교육 - 약초 판별 가이드 북 활용 권장 - 구분(판별)이 어려운 경우 전문가 동행 촬영 및 전문가 자문 필수 ② 촬영 교육 및 세미나 - 선정 약초 촬영 문제점 및 개선 방안 토의 - 촬영 노하우 공유 - 안전 사고 예방 및 신속 조치 - 저작권 등 법적 절차 동의 서류 확보 ③ 촬영 전 방문지 협조공문 발송 - 전국 농업기술원, 원예특작과학원, 식물원 등 ④ 약초 전문가 동행 촬영 일정 수립 ⑤ 약초 촬영 및 지원 - 수집 목표 관리 - 촬영 일정 계획 수립(선정 약초 보유 기관 정보 탐색 등) - 선정 약초 위치 정보 공유 - 촬영 매뉴얼 배포 - 약초 판별 가이드 북 배포(별첨 자료) ⑥ 명확한 구분이 어려운 경우 오픈 채팅을 통해 확인 후 촬영 - 카톡 단체 채팅방 개설 및 운영 - 실시간 질의응답 체계 - 국내 최고 약초 전문가 실시간 응답 ⑦ 촬영 이미지 분류 - 약초명, 촬영 장소, 촬영 부위 ⑧ 허브넷 등록 ⑨ 임무 종료

구분	내용
데이터 정제	① 기술적 정제(검수) - 촬영 이미지 기술적, 기계적 정제(검수) - 포커스, 흔들림, 겹침 등 정제 기준 적용 - 중복 등록 처리(허브넷 프로그램을 통해 자동 삭제 처리) ② 전문가 정제(검수) - 독성이 있는 약초 또는 비교식물 집중 정제(검수) - 카테고리 이동 - 강제 삭제 처리 ③ 임무 종료 ④ 정제 데이터는 데이터 가공에서 다운 받을 수 있도록 API 제공

2.4 획득 및 정제 기준

- 약초(비교식물 포함) 이미지 촬영 기준(공통 적용)
  - 선정 촬영 매뉴얼 준수 및 1개체 당 촬영 목표 제시

〈표 Ⅲ-127〉 약초 이미지 촬영 방법(예시)

구분	내용
전초(지상부) 촬영 방법	 

구분	내용
전초(지상부) 촬영 방법	 
줄기 촬영 방법	
잎 촬영 방법	

구분	내용
꽃 촬영 방법	 
열매 촬영 방법	
뿌리 촬영 방법	
뿌리 촬영 방법	

● 약초(비교식물 포함) 이미지 촬영 예시

〈표 Ⅲ-128〉 약초 이미지 촬영 예시

촬영 부위	촬영 방법	예 시
전초	대상 식물이 잘리지 않게 하고 화면에 70%가 채워지도록 촬영	
잎	싱싱한 잎, 어린잎, 잎의 크기가 다양한 것을 한 화면에 담기도록하고 잎은 구분이 명확한 잎을 위주로 화면에 70%가 채워지도록 촬영	
줄기	대상 식물의 특성을 알 수 있도록 가지 또는 잎이 나는 부위를 포함하여 1마디, 2마디를 화면에 70%가 채워지도록 촬영	
열매	열매의 특성이 나타나도록 근접 촬영하되 화면에 70%가 채워지도록 촬영, 포커스가 흔들리지 않도록 주의	
꽃	대상 식물의 꽃이 화면에 70%가 채워지도록 촬영 꽃이 여러개 일 경우에는 여러개의 꽃을 한 화면에 가득 채워지게 촬영	 
뿌리	뿌리의 형태를 볼 수 있도록 화면에 70%가 채워지도록 촬영, 근접 촬영 시 포커스 주의	

3 어노테이션/라벨링

3.1 어노테이션 / 라벨링 절차

[그림 Ⅲ-163] 데이터 가공 절차

- 저작 도구를 사용하여 정제 완료된 원시 데이터 중 전초, 잎, 꽃, 열매를 대상으로 바운딩 박스후 크롭 이미지 생성
- 원시 데이터 1장에서 크롭 이미지는 1장 이상 라벨링 가능
- 생성된 크롭 이미지에 어노테이션 정보(메타 데이터)를 추가하여 라벨링 완료
- 예시, 갈근(칡)

〈표 Ⅲ-129〉 원시데이터 Cropping 및 저장 예시

적 용	부위	크롭추출	크롭핑	어노테이션	저장	AI 가치
원시 데이터 (정제 완료)	전초	전초		메타데이터	json	높음
		잎		메타데이터	json	매우 높음
		꽃		메타데이터	json	매우 높음
		열매		메타데이터	json	높음

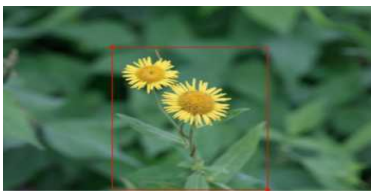
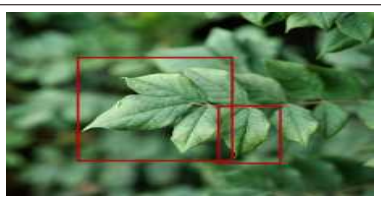
적 용	부위	크롭추출	크롭핑	어노테이션	저장	AI 가치
		줄기		메타데이터	json	낮음
		뿌리		메타데이터	Json	매우 낮음
	잎	잎		메타데이터	json	매우 높음
		꽃		메타데이터	json	매우 높음
	꽃	꽃		메타데이터	json	매우 높음
		잎		메타데이터	json	매우 높음
	열매	열매		메타데이터	json	높음
		잎		메타데이터	json	매우 높음

3.2 어노테이션 / 라벨링 기준

- 일반적인 기준

〈표 Ⅲ-130〉 약초 데이터 어노테이션 및 라벨링 기준

기준	세부 내용
공통	• 약초의 특징 정보를 잘 표현하고 있는 선명한 이미지를 선택해야 함 • 한 사진에서 여러 부위를 라벨링 할 때에는 지정해줄 약초부위명에 유의해야 함 • 기본 설정에서는 연관 이미지가 바이너리 형태로 변환되어 json 파일 내부에 대용량으로 저장되므로 Save With Image Data 옵션을 반드시 체크를 해제해야 함 • 폴더의 작업을 마쳤으면 작업한 jpg와 생성된 json 파일 개수가 일치하는지 확인 • 바운딩 박스가 객체의 안쪽이나 바깥쪽이 아닌 경계에 위치해야 함

<table>
<tr><th>기준</th><th>세부 내용</th></tr>
<tr><td></td><td>
• 바운딩 박스와 객체간의 거리는 상하좌우 100 픽셀 이내에서 위치 가능함
• 객체가 잘려서 촬영된 때에는 이미지 가장자리까지 사각형을 그리도록 함
• 라벨링(크롭)되는 객체에 다른 객체가 포함되면 안됨
<table>
<tr><td>  </td><td>  </td></tr>
<tr><td>X</td><td>X</td></tr>
<tr><td>  </td><td>  </td></tr>
<tr><td>O</td><td>O</td></tr>
<tr><td>  </td><td>  </td></tr>
<tr><td>O</td><td>O</td></tr>
</table>


• 동의보감 약초 판별 AI 데이터 라벨링 종류의 가공 대상은 모두 라벨링함을 원칙으로 하되, 아래의 경우에는 어노테이션 하지 않음
- 초점이 맞지 않아 객체가 흐려보이는 경우
- 객체가 지나치게 어둡거나 밝은 경우
- 그림자로 인하여 명암 대비가 확연히 일어난 경우
- 이미지 대비 약 10% 이하의 크기를 지닌 객체
</td></tr>
<tr><td>꽃</td><td>
• 기본적으로는 군집을 이룰 때에는 객체 하나씩 모두 사각형을 그려야 하지만, 작은 꽃잎들이 모인 군집이 하나의 꽃인 형태일 때는 군집으로 라벨링 할 수 있음
• 가려져 있는 꽃에 대해서는 라벨링을 실시하지 않음
</td></tr>
</table>

기준	세부 내용
잎	• 가장 전면에 나와있는 포커스가 정확한 잎을 작업 요망 • 병든 잎, 시든 잎 등은 라벨링 하지 않음. 다만, 원천 이미지의 품질이 좋고 병들거나 시든 정도가 약한 경우는 라벨링 할 수 있음(1차 검수에서 필터링 예정) • 가려져 있는 잎에 대해서는 어노테이션을 실시하지 않음 • 라벨링 시에 잎의 개수는 무방하게 작업을 진행하면 되나, 한 이미지당 최대 20개까지만 라벨링을 하도록 작업 요망
불인정 예시	 혼동  병든 잎  명암 대비  어두움  포커스불량

3.3 어노테이션 / 라벨링 교육

- 관련 내용 없음

3.4 어노테이션 / 라벨링 도구 및 사용법

1) 저작도구 : LabelMe

① 다운로드 사이트 : https://github.com/wkentaro/labelme

② Python 사용자를 위한 요구사항

a. Anaconda / pip

b. Python3

- Anaconda Prompt 나 Terminal에서 아래의 코드를 실행

```
# python3
conda create --name=labelme python=3.6
conda activate labelme
conda install pyqt
conda install labelme -c conda-forge
```

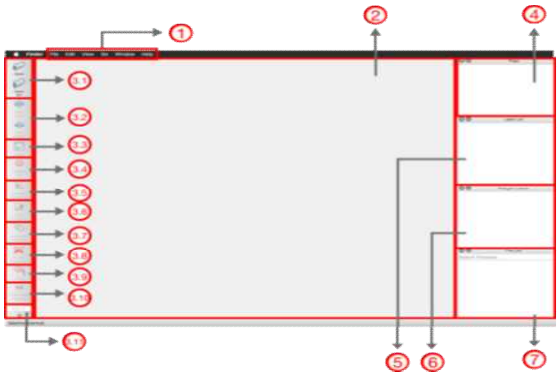
Python

2) 저작도구 실행

a. 'labelme.exe'를 더블클릭하거나, Anaconda Prompt나 Terminal에서 'labelme'를 입력/실행

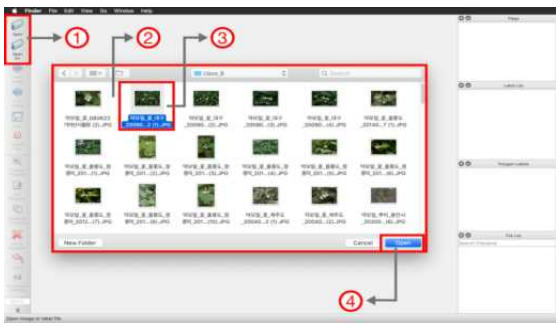
b. 실행된 화면의 구조는 아래와 같다.

〈표 Ⅲ-131〉 저작도구 LabelMe 실행화면 설명

화면 이미지	내용
 〈LabelMe 실행화면 구조〉	① 메뉴 영역 ② 메인 이미지 확인 영역 ③ 도구 화면 3.1) 필요한 이미지를 불러오기 - Open 이미지 하나만 불러오기 - Open Dir 이미지 폴더 불러오기 3.2) 현재 이미지 폴더 안에 다음/이전 이미지를 불러오기 3.3) 저장하기 3.4) 이미지 삭제 3.5) 사각형 어노테이션 그리기 3.6) 어노테이션 편집 3.7) 어노테이션 중복 3.8) 어노테이션 삭제 3.9) 되돌아가기 3.10) 밝기/콘트라스트 조절 3.11) 이미지 크기 변경 ④ 플래그 화면 ⑤ 라벨 리스트 화면 ⑥ 폴리곤 라벨들 화면 ⑦ 파일 리스트

- Step1 : 이미지 불러오기

〈표 Ⅲ-132〉 저작도구 LabelMe: 이미지 불러오기 방법

화면 이미지	내용
	① 작업 이미지를 불러오기 - Open : 이미지 하나만 불러오기 - Open Dir : 이미지 폴더 불러오기 ② 새로운 창이 열림 ③ 작업 이미지/폴더 선택하기 ④ 파일 선택 후 이미지 불러오기

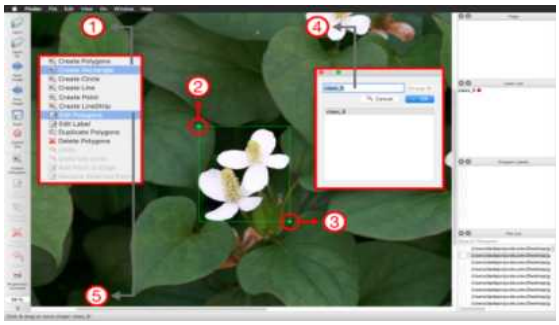
- Step2 : 이미지 확인

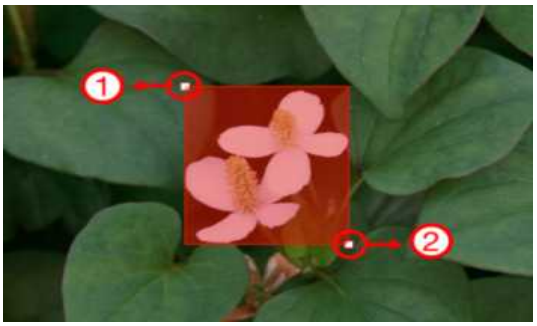
〈표 Ⅲ-133〉 저작도구 LabelMe: 이미지 확인 방법

화면 이미지	내용
	① 이미지 확인 - 작업 이미지가 정상적으로 노출되는지 확인 ② 이미지 폴더 확인 - 작업에 필요한 폴더가 정상적으로 노출되는지 확인

- Step3 : 드로잉

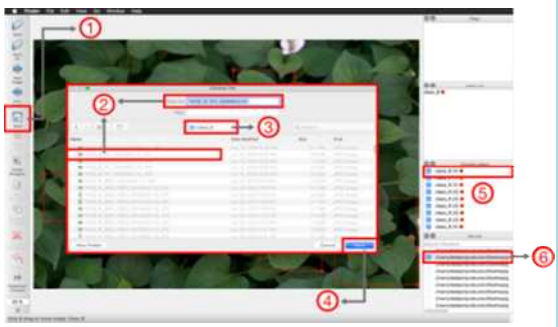
〈표 Ⅲ-134〉 저작도구 LabelMe: 드로잉 방법

화면 이미지	내용
	① 이미지 위에서 마우스 우클릭 후 새로 열린 창에서 '사각형 그리기(create rectangle)' 클릭하여 라벨링 진행 시작 ② 드로잉 1 - 첫 번째 점 드로잉 - 초록색 실선으로 된 가이드 선을 참고하여 대상 외곽 기준 확인 ※ 라벨링을 잘못하였을 때 : Delete 키를 눌러 드로잉 제거 ③ 드로잉 2 - 두 번째 점 드로잉 - 초록색 실선으로 된 가이드 선 참고하여 대상 외곽 기준 확인 ④ 라벨 선언 - 두 번째 점을 드로잉하면 자동으로 새로운 창이 열려 라벨링 대상(약초부위명) 입력 - '확인'(ok) 클릭하고 라벨 선언 ※ 라벨링을 편집 - 이미지 위에 우클릭 후 '폴리곤 편집' (edit polygon) 선택 - ②, ③번을 클릭 후 드래그 하여 편집

화면 이미지	내용
	예) ①, ② 점 중 수정이 필요한 점 위에 마우스 이동하여 하얀색 네모로 변경된 상태에서 마우스 왼쪽 클릭 후 드래그 하면서 수정 작업 진행
	⑤ 라벨링 결과 확인 - 라벨링 작업을 수행 한 후 저작도구(Labelme) 화면에서 확인할 수 있는 가공작업

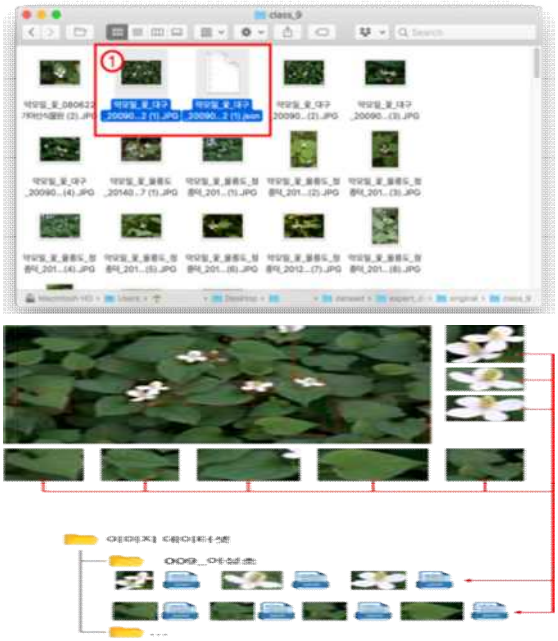
- Step4 : 저장하기

〈표 Ⅲ-135〉 저작도구 LabelMe: 저장 방법

화면 이미지	내용
	① 저장하기 - 드로잉과 라벨 값 입력 후 저장 - 저장하기 (Save) 버튼으로 진행하며 이미지 한 장마다 해당 버튼으로 파일을 저장 ② 저작도구 작업 후 원본 이미지와 파일명이 동일해야 함 - 자동으로 동일한 파일명이 정해빔 ③ 원본 이미지와 같은 폴더에 저장 ④ 저장하기 버튼을 눌러 저장 ⑤ 라벨 선언 및 저장 확인 - 저장 여부 확인 가능 - 저장 전 : 체크 안되어 있음 - 저장 후 : 체크

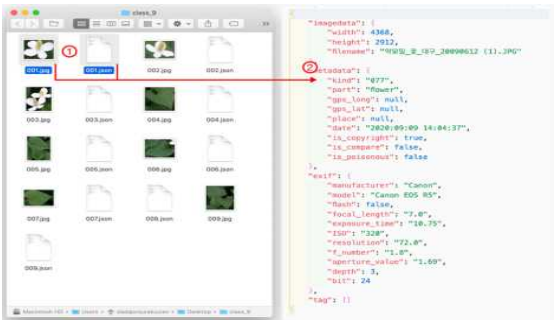
- Step5 : 크롭핑(Cropping)

〈표 Ⅲ-136〉 저작도구 LabelMe: 크롭핑 방법

화면 이미지	내용
	① Step4 완료 후 해당 폴더에 원천 이미지 데이터 파일 이름과 동일한 JSON 파일이 생성되었음을 확인할 수 있음(①) ② 이전 단계(Step4)에서 생성된 MS COCO 데이터셋의 JSON 파일을 기반으로 원천 이미지에서 인스턴스들을 자르고(크롭, Crop) 새로운 이미지 파일을 만듦(내부 가공) - 원천 이미지 데이터에서 인공지능 학습용 이미지 데이터를 여러 개 생성할 수 있음

- Step6 : 어노테이션(Annotation)

〈표 Ⅲ-137〉 저작도구 LabelMe: 어노테이션 방법

화면 이미지	내용
	① 1차 데이터 라벨링(크롭핑) 과정에서 학습용 이미지 데이타 파일마다 데이터 속성들을 포함하는 메타데이터들을 추가함(내부 가공) ② 데이터 라벨링이 이루어진 약초명(코드 id)으로 새로운 폴더를 만들고 이미지 파일과 최종 메타데이터 파일(JSON)을 해당 폴더에 저장함 ③ 결과물 확인 - 이미지가 있는 폴더로 이동하여 해당 이미지에 대한 결과물 확인 - 메타 데이터가 포함된 JSON 파일 생성 확인 ④ 메타 데이터 확인 - 해당 파일을 실행하여 메타 데이터 구조 및 값 확인

4 데이터 검수

4.1 검수 절차

● 라벨러 작업자 교차 검수 절차

〈표 Ⅲ-138〉 라벨러 작업자 교차 검수 방법

절 차	• 라벨링 → 라벨링 참여자 간 교차 확인 → 부적합 시 : 검수자 정정
조 직	• 라벨링 참여자 및 검수 관리자
검수도구	• 라벨링 과정에서 사용한 저작도구(Labelme)

● 라벨링(크롭핑) 작업 후 검수 절차

〈표 Ⅲ-139〉 라벨링 작업 후 검수 방법

1차 검수 (일반 검수)	절 차	• 학습용 이미지 데이터 중복성 확인 → 중복 데이터 제거 • 학습용 이미지 데이터 확인 → 부적합 시 : 저장(부적합 데이터) • 생성된 JSON 파일 확인 → 구문 오류시 : 수정후 저장
	조 직	• 내부조직 및 아웃소싱
	검수도구	• 중복성 체크 프로그램 • 이미지 뷰어(jpg 파일을 열어 이미지를 확인할 수 있는 툴) • 메타데이타 검수 프로그램(json schema Validator)
2차 검수 (전문 검수)	절 차	• 1차 검수된 이미지 데이터 확인 → 적합 시 : 저장(학습 데이터) → 부적합 시 : 저장(부적합 데이터)
	조 직	• 인공지능 학습모델 담당자/전문가 • 한의학 또는 식물분류 전문가
	검수도구	• 이미지 뷰어(jpg 파일을 열어 이미지를 확인할 수 있는 툴)

4.2 검수 기준

● 라벨러 작업자 교차 검수 기준

〈표 Ⅲ-140〉 라벨링 작업자 교차 검수 기준 제시

기준 및 방법	• 라벨링 작업 오류 검수 • 2인이상 교차검증
상세기준	• 표기 오류, 오타 오류, 라벨링 위치 잘못 지정 등 검수 • 작업자의 실수로 인해 박스 위치가 잘못 지정된 데이터 선별 • 클래스 표기 오류(예, flower를 leaf로 잘못 표기) 데이터 선별

● 라벨링(크롭핑) 작업 후 검수의 기준

〈표 Ⅲ-141〉 라벨링 작업 후 검수의 기준 제시

1차 검수 (일반 검수)	기준 및 방법	• 학습용 이미지 데이터 품질 검수(중복성, 품질) • 전수검수
	상세기준	• 학습용 이미지 데이터가 중복된 경우 • 이미지내 식물의 차지 비율이 작은 경우 • 선명하지 않거나 초점이 흐린 경우 • 학습용 이미지로 품질이 떨어지는 경우 • JSON 파일 포맷의 속성값의 오류가 입력된 경우
2차 검수 (전문 검수)	기준 및 방법	• 인공지능 학습에 적절한지, 약초분류가 제대로 되어 있는지 • 전수검수
	상세기준	• 인공지능 학습측면 : 학습용 이미지로서의 품질 검수 • 한의학/식물분류측면 : 이미지 데이타의 식물명 오류 최종 검수 • 전문 검수에서 이견이 발생 할 경우 다수 전문가들이 참여하여 다수결 원칙에 따라 적합/부적합 검수

5 데이터 활용 방안

5.1 학습 모델

● 학습 모델 선정 근거

〈표 Ⅲ-142〉 학습 모델 선정 근거

연구내용	
Evidence 1 • Grant Van Horn, Oisin Mac Aodha et al. • CVPR 2018. • The iNaturalist Species Classification and Detection Dataset Bounding boxes were not collected on the Plantae, Fungi, Protozoa or Chromista super-classes because these super-classes exhibit properties that make it difficult to box the individual instances (*e.g.* close up of trees, bushes, kelp, *etc.*). An alternate form of pixel annotations, potentially from a more specialized group of crowd workers, may be more appropriate for these classes.	

	Super-Class	Class	Train	Val	BBoxes
	Plantae	2,101	158,407	38,206	-
	Insecta	1,021	100,479	18,076	125,679
	Aves	964	214,295	21,226	311,669
	Reptilia	289	35,201	5,680	42,351
	Mammalia	186	29,333	3,490	35,222
	Fungi	121	5,826	1,780	-
	Amphibia	115	15,318	2,385	18,281
	Mollusca	93	7,536	1,841	10,821
	Animalia	77	5,228	1,362	8,536
	Arachnida	56	4,873	1,086	5,826
	Actinopterygii	53	1,982	637	3,382
	Chromista	9	398	144	-
	Protozoa	4	308	73	-
	Total	5,089	579,184	95,986	561,767

연구내용

Evidence 2

- Subhransu Maji, Esa Rahtu, Juho Kannala et al
- CVPR 2013
- Fine-Grained Visual Classification of Aircraft

FGVC-Aircraft contains 100 example images for each of the 100 model variants. The image resolution is about 1-2 Mpixels. Image quality varies as images were captured in a span of decades, but it is usually very good. The dominant aircraft is generally well centred, which helps focusing on fine-grained discrimination rather than object detection. Images are equally divided into training, validation, and test subsets, so that each subset contains either 33 or 34 images for each variant. Algorithms should be designed on the training and validation subsets, and tested just once on the test subset to avoid over fitting.

Evidence #

- Jiongxin Liu , Angjoo Kanazawa et al
- ECCV 2012
- Dog Breed Classification Using Part Localization

Abstract. We propose a novel approach to fine-grained image classification in which instances from different classes share common parts but have wide variation in shape and appearance. We use dog breed identification as a test case to show that extracting corresponding parts improves classification performance. This domain is especially challenging since the appearance of corresponding parts can vary dramatically, e.g., the faces of bulldogs and beagles are very different. To find accurate correspondences, we build exemplar-based geometric and appearance models of dog breeds and their face parts. Part correspondence allows us to extract and compare descriptors in like image locations. Our approach also features a hierarchy of parts (e.g., face and eyes) and breed-specific part localization. We achieve 67% recognition rate on a large real-world dataset including 133 dog breeds and 8,351 images, and experimental results show that accurate part localization significantly increases classification performance compared to state-of-the-art approaches.

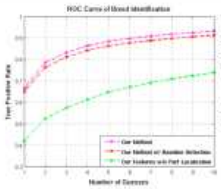
Evidence #

- Peter N. Belhumeur, Daozheng Chen et al
- ECCV 2008
- Searching the World's Herbaria: A System for Visual Identification of Plant Species

Abstract. We describe a working computer vision system that aids in the identification of plant species. A user photographs an isolated leaf on a blank background, and the system extracts the leaf shape and matches it to the shape of leaves of known species. In a few seconds, the system displays the top matching species, along with textual descriptions and additional images. This system is currently in use by botanists at the Smithsonian Institution National Museum of Natural History. The primary contributions of this paper are: a description of a working computer vision system and its user interface for an important new application area; the introduction of three new datasets containing thousands of single leaf images, each labeled by species and verified by botanists at the US National Herbarium; recognition results for two of the three leaf datasets; and descriptions throughout of practical lessons learned in constructing this system.

Evidence #

- Jiongxin Liu , Angjoo Kanazawa et al
- ECCV 2012
- Dog Breed Classification Using Part Localization



The first uses our feature set sampled on a grid within our detected face window. The second uses part localization, but applied to only the highest scoring window from the face detector. The third – and best – uses both the part scores and face detection scores to select the best face window.

- 위 논문을 근거로 이 과제에서 정확도/시간/효율성 등을 고려하여 아래와 같은 결론 도출

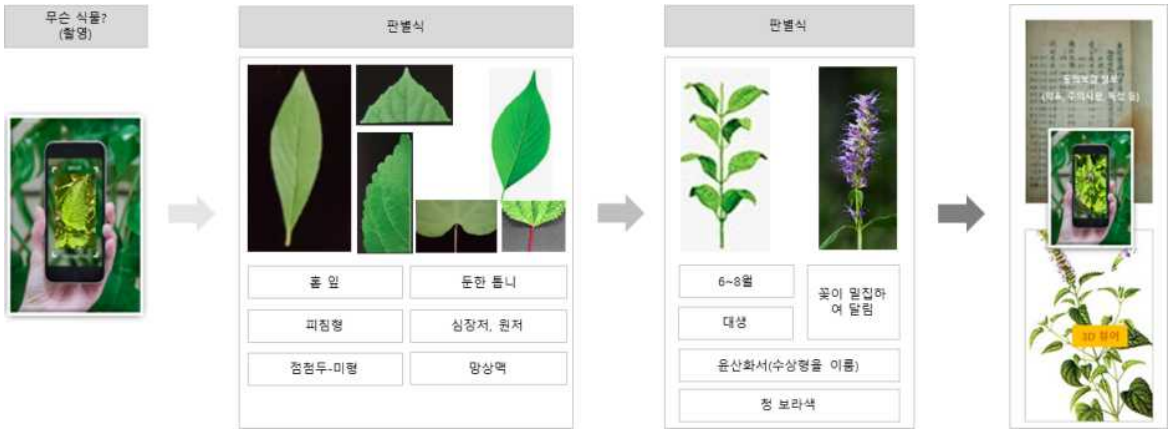
● 학습 모델 알고리즘

- 목적 : 분류
- 학습데이터 : 판별 안된 라벨링 작업이 완료된 식물 이미지 사진(JPG)
- 훈련, 검증, 테스트 데이터의 비율 : 8:1:1
- AI 학습모델 계열 및 적용 가능 AI 학습모델 : CNN 계열, DenseNet169
- 추론(예측)

※ 입력데이터 : 판별 안된 식물 이미지 사진(JPG)
※ 출력데이터 : 클래스별 F1-Score

5.2 서비스 활용 시나리오

- 응용 서비스(사용자의 직관적 시각효과를 높일 수 있도록 어플리케이션에 3D 뷰어 적용)



[그림 Ⅲ-164] 응용 서비스 예시

- 선정약초 50종에 대한 3D 모델링 제작 및 검증

[그림 Ⅲ-165] 약초 3D 모델링 검증

제14장 CCTV 영상 AI 데이터

1 데이터 정보 요약

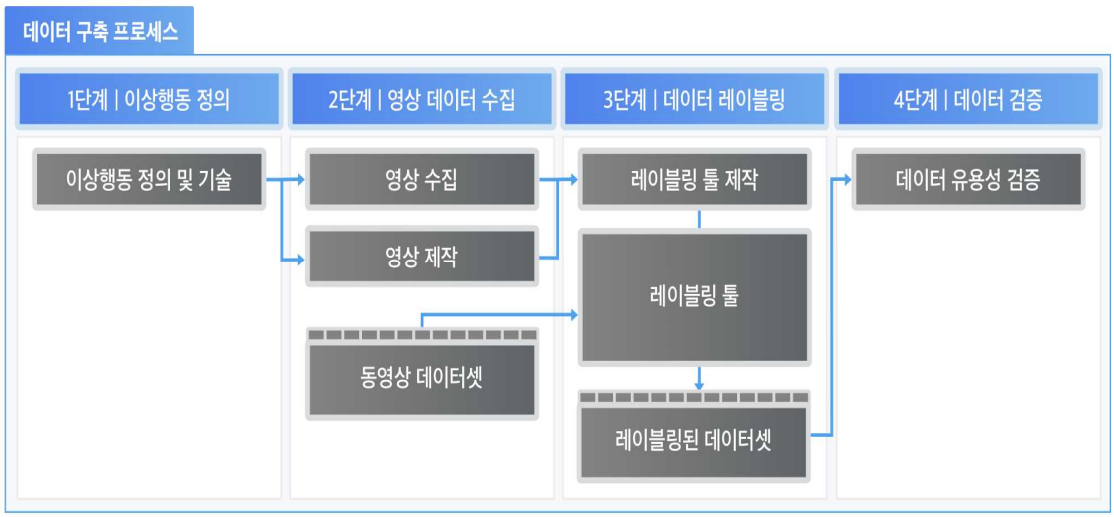
1.1 가이드 분류

대분류	이미지	중분류	비전	소분류	JPG

1.2 데이터 정보

데이터 이름	CCTV 영상 AI 데이터
활용 분야	CCTV 이상행동 및 객체 추적, 범죄 예방, 교통약자 돌봄 및 안전사고 예방
데이터 요약	도시철도 역사 내 CCTV에서 관측될 수 있는 이상행동 영상 데이터 및 사회적 약자 보호, 범죄 예방을 위한 객체 추적 영상 데이터
데이터 출처	• CCTV 영상 자체 촬영 및 편집, 크로마키 배경 촬영 • 대전 도시철도공사 기 보유중인 CCTV 영상 활용

1.3 데이터 구축 개요



[그림 Ⅲ-166] CCTV 영상 데이터 구축 프로세스

- 이상 행동 정의
  - CCTV 영상에서 사람 혹은 여러 명의 이상행동 패턴을 구체적으로 정의하는 단계로 도시철도 역내의 실제 상황을 반영한 이상행동 패턴 정의를 위해 대전 도시철도의 CCTV 영상을 모니터링 후, 행동 패턴을 정의
- 영상 데이터 수집 및 정제
  - 정의한 이상행동을 포함하는 CCTV 영상을 수집 또는 제작하는 단계로 본 과제에서는 ① CCTV 영상 자체 제작 ② 대전 도시철도공사가 기 보유중인 CCTV 영상 활용 및 신규 수집되는 영상을 이용
  - 데이터 정제는 기존 영상중 사용 할 수 없는 영상을 제거하고(이상행동의 식별이 어렵거나, 화질이 떨어지는 영상 등) 개인정보 비식별화 작업을 수행(얼굴 모자이크 처리 등)
- (레이블링)
  - 수집 또는 촬영한 영상을 크라우드 소싱을 통해 프레임별로 객체 정보 및 이상행동 정보를 레이블링하는 단계
- (데이터 검증)
  - 레이블링된 영상들을 최신 인공지능 모델을 사용하여 학습하고 그 성능을 평가하여 데이터셋과 레이블링된 정보의 유용성을 검증하는 단계
  - 각 소주제별로 학습용 데이터와 검증용 데이터의 비율은 4:1로 함
  - 검증 단계에서 영상에 대한 비식별화 확인 작업을 별도로 수행하며, 비식별화가 되지 않은 영상 발견 시 2차 정제 작업 수행

1.4 구축 목적

- 도시철도 내 CCTV를 이용한 이상행동 감지, 이상행동 판별, 동일인물 추적을 통한 승객 교통안전 증대, 사회범죄 예방 및 교통약자 보호를 위한 AI 학습용 영상 데이터 구축, AI 학습 모델 제시 및 테스트베드 구축에 의한 유효성 검증
- 안전하고 편리한 도시철도 역사 실현 및 AI 연구분야 활용, 사업화 및 판로 개척

1.5 활용 분야

- CCTV 이상행동 및 객체 추적, 범죄 예방, 교통약자 돌봄 및 안전사고 예방

1.6 유의 사항

- 기확보된 데이터는 비식별화를 위해 라벨링 작업 이전에 안면인식 알고리즘을 활용하여 사람 얼굴에 대한 모자이크 처리
- 데이터 검증 단계에서 알고리즘 적용 후 누락 된 안면에 한하여 육안으로 확인된 데이터는 비식별화 작업 수행
- 시나리오 촬영 시에는 배우에게 동의서 수령

2 데이터 획득 및 정제

2.1 원시 데이터 선정

● 원시 데이터 선정 절차

- 실내 환경(건물 내부 및 도시철도역사 등)에서 발생할 수 있는 이상행동과 추적이 필요한 객체를 정의함
- 이상행동의 경우 실제 발생 빈도 고려하고, 객체 추적의 경우 교통 약자 및 범죄 예방에 직접적인 연관이 있는 클래스를 선정
- 대전 도시철도공사에서 기 보유중인 영상 중, 화질이 FHD이상(1920*1080)인 영상을 별도로 제공받아서 검증 데이터로서 활용하며, 그 외의 영상은 모두 전문 촬영업체를 통해 시나리오 기반으로 촬영함
- 이상행동 영상의 데이터셋에서 정상행동과 이상행동의 비율은 4:1(400시간 정상행동, 100시간 이상행동)
- 시나리오 촬영 방안은 다음과 같음

〈표 Ⅲ-143〉 시나리오촬영 방안

촬영 환경 조건		대응 방안
계절성		자체제작 영상의 경우 다양한 계절성을 반영하기 위하여 계절별 복장을 준비하여 촬영
시간대	주간	주간 조명 환경으로 촬영
	야간	야간 조명 환경으로 촬영
카메라 해상도		4K 해상도로 촬영

촬영 환경 조건	대응 방안
배경	도시철도역사 내 다양한 공간에서 촬영
	에스컬레이터 전도 영상획득을 위해 에스컬레이터 공간에서 촬영
	개집표기 이상행동 영상 획득을 위해 개집표기가 존재하는 공간에서 촬영 데이터 획득

- 도시철도 역사 내 시민들의 원활한 이용을 위해 야간 조명 촬영이 어려우며, 시간대를 확인할 수 있는 영상의 경우 외부에 있는 에스컬레이터에서 촬영한 영상뿐이며, 해당 장소에서 촬영한 데이터 역시 낮에 촬영되었기 때문에 야간 환경으로 인식될 만한 데이터는 없음외부에서 촬영된 데이터는 전체 데이터의 약 0.8%임.
- 피촬영자는 촬영 업체에서 섭외한 전문 배우들로 구성하여, 나이, 성별로 균등하게 배분하여 이상행동 및 객체 추적 영상 촬영
- 촬영 대상자에게는 반드시 개인정보 활용 동의서를 작성하여 법적 절차를 거친 후에 촬영 및 데이터 공개하며, 기 보유중인 영상 데이터의 경우 비식별화 작업을 거친 후 공개에 대한 별도 법률 자문을 구할 예정

● 이상행동 영상 데이터

- 이상행동 / 상황 정의
- - 각 도시철도 역에서 발생 빈도수가 가장 많은 이상행동을 다음과 같이 정의함

〈표 Ⅲ-144〉 이상행동 행동명 및 내용

이상 행동		묘사
번호	행동명	
①	에스컬레이터 전도	• 난간 손잡이를 잡지 않고 E/S의 속도에 의해 중심을 잃고 전도 • 자전거, 짐수레 등에 의한 전도 • 건강상의 이유(빈혈등)로 정신을 잃고 전도
②	계단전도	• 핸드폰 등 또는 뛰어 오르다가 헛디뎌 중심을 잃고 전도
③	환경 전도	• 바닥 물기에 의한 미끄러짐 • 개집표의 부주의로 게이트에 걸려 전도 • 엘리베이터 급정거에 의한 전도
④	몰카촬영	• E/S, E/V,승강장등 도시철도역 전구간에서 간헐적으로 핸드폰 등 전자기기를 이용하여 몰래카메라 촬영
⑤	주취행동	• 음주후 E/S 이용시 주저앉거나 전도 • 대합실, 승강장 등 쉼터 또는 벤치 등에서 잠을 자거나 비틀거리며 걸어감 • 역구내에서 지나가는 사람에게 시비를 걸거나 폭행

이상 행동		묘사
번호	행동명	
⑥	배회	• 치매로 인하여 자신의 위치를 정확히 인지하지 못하고 주변을 두리번거리며 배회 • 성추행 등의 전조증상으로 범행 장소에서 또는 범행을 하고자 하는 사람을 찾으며 배회
⑦	실신	• 심정지, 간질, 과호흡 등 건강상의 문제로 인하여 실신
⑧	기물파손 (스크린 도어)	• 역구내 공공시설물을 파손하는 행위
⑨	유기	• 테러(폭파, 유독가스)를 위하여 역구내에 물건을 유기 • 역구내에서 대기 중 물건을 놓고 이동
⑩	폭행	• 안면이 있는 사람끼리 의견충돌 또는 다양한 이유에 의하여 단순폭행 • 연인끼리 데이트 중 서로의 의견 차이로 폭행
⑪	절도, 강도	• 역구내에서 핸드백, 손가방, 핸드폰 등을 절도
⑫	개집표기 출입방향 오인	• 개집표기 출입 방향과 반대 방향으로 진입 시도 • 출입방향 오인으로 인한 진입 실패
⑬	개집표기 무단진입	• 개집표기를 뛰어 넘어 무단 진입

● 객체 추적 영상 데이터

〈표 Ⅲ-145〉 객체 추적 영상 행동명 및 묘사

피추적자		묘사
번호	행동명	
①	휠체어 이용자	• 수동, 전동 휠체어 사용자
②	시각 장애인	• 시각장애인 보조 도구 사용자
③	유모차 이용자	• 아이를 동반한 유모차 이용자
④	주취자	• 비틀거림, 주저앉음 등
⑤	잡상인	• 금지된 물품 판매 행위자(카트, 대형 배낭 등 소지자)
⑥	아동	• 12세 미만 아동

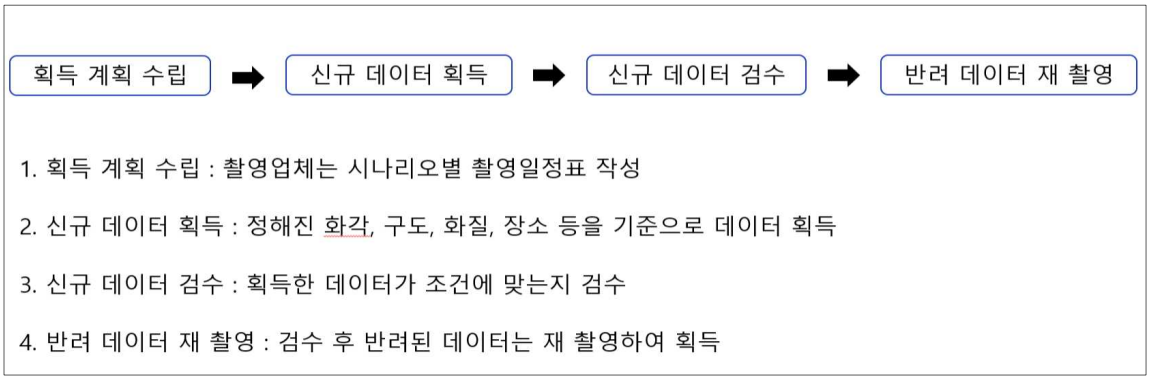
2.2 규제관련 사항

- 관련 내용 없음

※ 본 데이터 가이드라인 상에는 규제 관련 내용이 없지만, 데이터 구축 시 규제 관련 내용 확인 및 검토 필요

2.3 획득 및 정제 절차

- 원천 데이터 획득 과정



[그림 Ⅲ-167] 원천 데이터 획득 절차

- 원천데이터 획득 시 촬영업체는 각 3팀씩 촬영팀을 구성하여 촬영일정표와 시나리오 작성 등 획득 계획을 수립하여 이행
- 원천데이터 획득 후 정제 과정에서 무의미하거나 불필요한 영상데이터 검출시 제거 후 재촬영을 진행하여 신규 데이터를 획득

2.4 획득 및 정제 기준

- 기 보유 영상
  - 해상도 : FHD(1920 * 1080), 30fps 이상
  - 이상행동을 명확히 확인 가능하거나, 추적 대상 객체가 존재하는 경우만 선정
  - 화질이 불량하거나, 렌즈 이물질 등으로 잘 보이지않는 영상의 경우 배제
- 신규 촬영 영상
  - 해상도 : 4k, 30 fps 이상
  - 화각 및 구도 : 대전 도시철도역사 기존 설치 CCTV 영상의 화각 및 구도 참조

- 촬영장소: 4개 대전 도시철도역에서 수행(시청역, 갑천역, 월드컵경기장역, 신흥역)
  ※ 촬영 상황에 따라 변경 가능
- 촬영 시나리오별 분량은 기존에 설정한 기준을 따르나, 기 보유 영상의 보유 비율에 따라 다소 조정될 수 있음
- 적절한 시나리오로 판별 될 수 있는지 여부 확인(다수결 원칙 활용)

3 어노테이션/라벨링

3.1 어노테이션 / 라벨링 절차

1) 데이터 형태
   - 원본 영상(혹은 초당 n개의 프레임으로 추출된 이미지)
   - 메타데이터 정보(날씨, 시간, 이상행동, 출연자 정보 등)
2) Step 1
   - 메타데이터에서 촬영 날씨, 시간 정보를 추출하여 각 동영상 파일명을 〈고유번호〉_〈날씨〉_〈시간〉_〈원본파일명〉.mp4 형식으로 변경하여 저장. 이때 고유번호는 중복되지 않도록 자동 생성
3) Step 2
   - 각 영상은 지정된 초당 프레임에 따라 추출하여, 동영상 파일명으로 생성된 폴더에 저장
   - 저작도구 관리자페이지에서 영상 라벨링 카테고리의 신규 프로젝트로 등록
   - 생성된 프로젝트 관리 페이지에서 원천 데이터 (프레임)를 불러오기 기능을 통하여 이미지 단위 작업물 생성
4) Step 3
   - 생성된 프로젝트 관리 페이지에서 참여할 작업자, 작업 단가, 주별 목표량 및 기타 프로젝트 세부설정을 등록
   - 작업자는 작업자용 저작도구에서 해당 프로젝트에 참여, 해당 시점에 특정 동영상에 해당되는 작업 묶음이 작업자에게 자동으로 일괄 배정됨

3.2 어노테이션 / 라벨링 기준

1) 개요 : 이상행동 구간 속에서 행동의 주체에 대해 바운딩 박스 형태의 라벨링을 하는 단계 또한 추적할 객체가 나타났을 때 객체에 대한 바운딩 박스 형태의 라벨링을 하는 단계

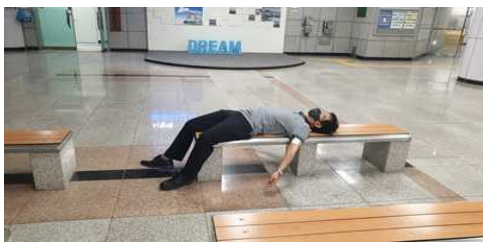
2) 작업 순서 및 방법

- 이상행동 구간의 객체는 '이상행동의 주체'에 대하여 라벨링 하며, 객체 추적 데이터의 경우 6가지 항목에 대하여 라벨링 함(휠체어, 시각장애인 등)
- Bounding Box는 라벨링 대상 객체의 전신(혹은 전 범위)를 정확히 포함하는 사각형 형태가 되어야 함
- Box 작업이 완료된 후에는 이상행동 혹은 객체 추적 대상에 대한 라벨링 작업을 수행함
- 모든 객체 라벨링이 완료되면 제출하기 버튼을 클릭하며, 해당 작업물은 '작업완료' 형태로 검수자에게 전달되며 작업자에게는 다음 작업할 이미지가 할당

● 이상행동 영상 데이터

〈표 Ⅲ-146〉 이상행동 영상 데이터 예시

이상 행동		묘사	사진
번호	레이블명		
①	escalator_fall	• 난간 손잡이를 잡지 않고 E/S의 속도에 의해 중심을 잃고 전도 • 자전거, 짐수레 등에 의한 전도 • 건강상의 이유(빈혈등)로 정신을 잃고 전도	
②	stairway_fall	• 핸드폰 등 또는 뛰어 오르다가 헛디뎌 중심을 잃고 전도	

이상 행동		묘사	사진
번호	레이블명		
③	surrounding_fall	• 바닥 물기에 의한 미끄러짐 • 개집표의 부주의로 게이트에 걸려 전도 • 엘리베이터 급정거에 의한 전도	
④	spy_camera	• E/S, E/V, 승강장 등 도시철도역 전 구간에서 간헐적으로 핸드폰 등 전자기기를 이용하여 몰래카메라 촬영	
⑤	public_intoxication	• 음주후 E/S 이용 시 주저앉거나 전도 • 대합실, 승강장 등 쉼터 또는 벤치 등에서 잠을 자거나 비틀거리며 걸어감 • 역구내에서 지나가는 사람에게 시비를 걸거나 폭행	
⑥	loitering	• 치매로 인하여 자신의 위치를 정확히 인지하지 못하고 주변을 두리번거리며 배회 • 성추행 등의 전조증상으로 범행 장소에서 또는 범행을 하고자 하는 사람을 찾으며 배회	
⑦	fainting	• 심정지, 간질, 과호흡 등 건강상의 문제로 인하여 실신	
⑧	property_damage	• 역구내 공공시설물을 파손하는 행위	

이상 행동		묘사	사진
번호	레이블명		
⑨	abandonment	• 테러(폭파, 유독가스)를 위하여 역구내에 물건을 유기 • 역구내에서 대기 중 물건을 놓고 이동	
⑩	violence	• 안면이 있는 사람끼리 의견충돌 또는 다양한 이유에 의하여 단순폭행 • 연인끼리 데이트 중 서로의 의견 차이로 폭행	
⑪	theft	• 역구내에서 핸드백, 손가방, 핸드폰 등을 절도	
⑫	turnstile_wrong_direction	• 개집표기의 출입방향(오른쪽에서 카드나 토큰을 테그)을 오인하여 개집표기 방향과 반대되는 방향으로 진입	
⑬	turnstile_trespassing	• 개표나 집표를 하지 않고 개집표기를 뛰어넘거나 플립이 닫혔음에도 무단으로 지나감	

● 객체추적 영상 데이터

〈표 Ⅲ-147〉 객체추적 영상 데이터 예시

피추적자		묘사	사진
번호	행동명		
①	wheelchair	• 수동, 전동 휠체어 사용자	
②	blind	• 시각장애인 보조 도구 사용자	
③	srollert	• 아이를 동반한 유모차 이용자	
④	drunk	• 비틀거림, 주저앉음 등	
⑤	merchant	• 금지된 물품 판매 행위자(카트, 대형 배낭 등 소지자)	

피추적자		묘사	사진
번호	행동명		
⑥	child	• 12세 미만 아동	

3.3 어노테이션 / 라벨링 교육

- 관련 내용 없음

3.4 어노테이션 / 라벨링 도구 및 사용법

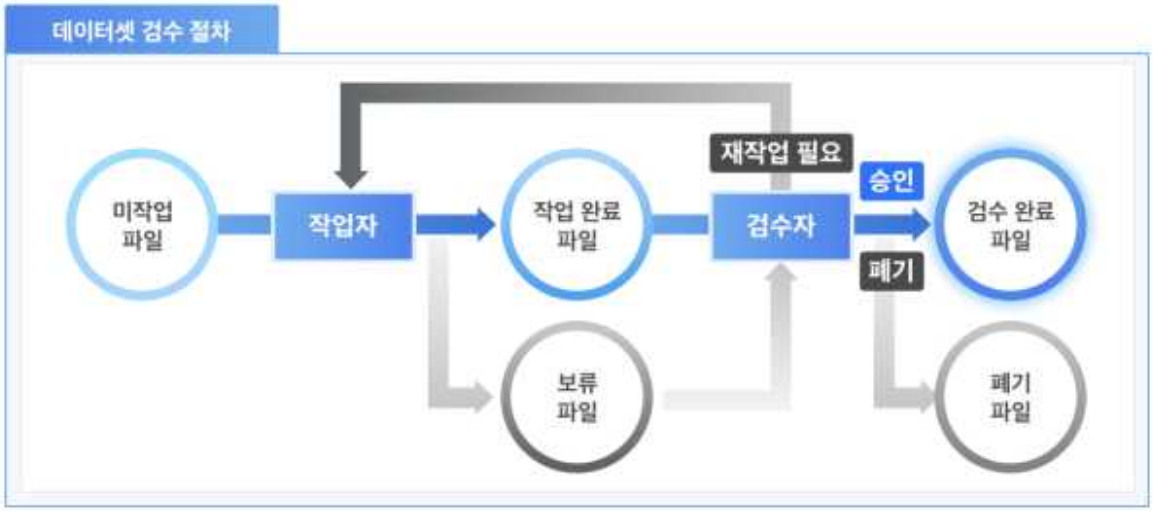
1) 개요: 영상 속에서 이상행동이 언제 발생하였는지, 언제 종료되었는지를 라벨링 하는 단계

2) 작업 순서 및 방법

- 이상행동의 기준은 어노테이션/라벨링 기준을 따름. 작업자는 해당 기준에 따라 라벨링을 수행
- 영상을 재생한 후 자유롭게 컨트롤 할 수 있으며, 스크롤이나 버튼 클릭 액션을 통하여 프레임 간 이동이 가능
- 영상 정지상태에서 버튼을 통해 이벤트 시작 프레임과 종료 프레임을 지정할 수 있으며 프레임 지정 후 적절한 카테고리의 이상행동을 선택

4 데이터 검수

4.1 검수 절차

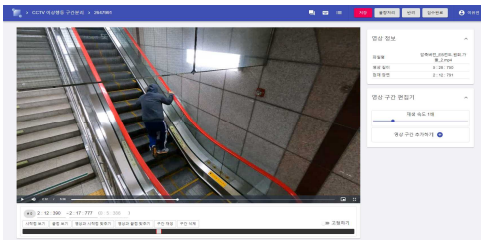
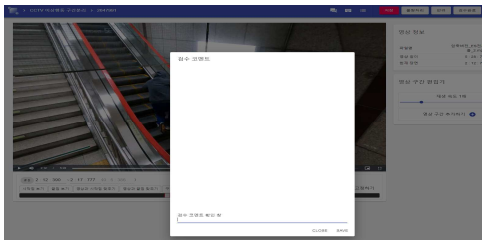


[그림 Ⅲ-168] 검수 절차

- 검수 절차 및 검수 방법
  - 작업물의 상태는 대기, 작업중, 작업완료, 검수완료(1차 검수), 승인(2차 검수), 반려, 보류로 구성
  - 대기: 작업물에 대해 작업자가 배정되지 않았거나, 배정되었으나 수행하지 않은 상태
  - 작업중: 배정된 작업물을 수행중인 상태
  - 작업완료: 작업자에 의해 작업이 완료된 상태
  - 검수완료: 1차 검수자에 의해 작업물이 통과된 상태
  - 승인: 2차 검수자에 의해 이중검수가 완료되고 포인트 지급이 가능한 상
  - 보류: 작업자 혹은 검수자가 판단하기 모호하거나, 작업 기준이 모호한 경우 선택하는 상태로 별도 판정단이 해당 작업물에 대해 판별한 후, 작업할 수 없다면 폐기처리, 작업할 수 있는 경우 가이드와 함께 반려처리함
  - 반려: 작업완료된 이후 프로세스에서 검수자에 의해 부적합 판정을 받은 작업물, 원 작업자에게 배정되어 재작업을 대기중인 상태

- 개별 작업상태에서 다음 검수 프로세스 진입 후, 부적합 판정 시 판정 사유와 함께 반려처리하여 작업을 수행한 작업자에게 재배정함

〈표 Ⅲ-148〉 검수 방법

검수 페이지	검수 코멘트 확인 창
	
검수자는 작업 Tool이 아닌 검수 페이지에 접속	검수 코멘트 창에 반려 사유 작성

4.2 검수 기준

- 이상행동 구간 라벨링
  - 선정한 이상행동의 분류는 올바른가?
  - 이상행동의 시작, 끝 구간의 지정은 올바른가?

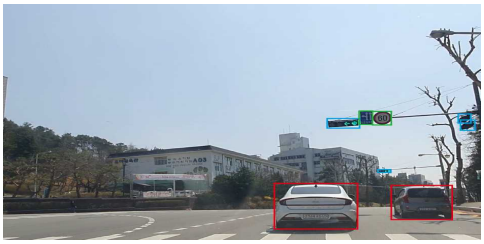
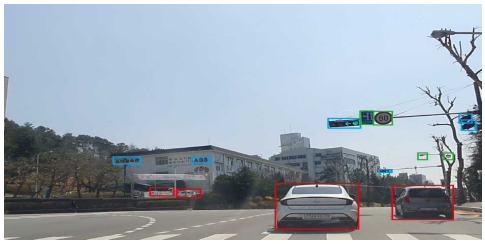
    ※ 시작, 끝 구간이 모호한 경우 3명 이상의 다수결 판단을 통해 확인 및 수정 진행
- 이상행동, 객체 추적 대상의 바운딩 박스 라벨링
  - 적절한 객체를 라벨링 하였는가?

    ※ 잘못된 객체를 라벨링한 경우(과검출),, 라벨링 해야할 객체를 누락한 경우(미검출) 반려

〈표 Ⅲ-149〉 미검출 반려 사유 예시

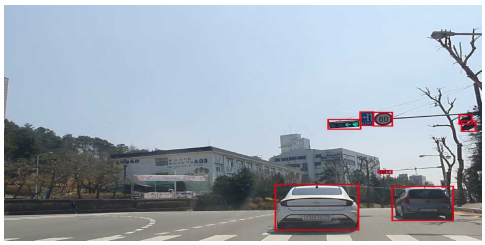
승인 예시	반려 예시(미검출)
	
	반려사유 : 흰색 차에 대한 라벨링 미시행)

〈표 Ⅲ-150〉 과검출 반려 사유 예시

승인 예시	반려 예시(과검출)
	
	반려사유 : 라벨링하지 않아야 할 차량라벨링

- 객체의 클래스는 올바른가?

〈표 Ⅲ-151〉 오검출 반려 사유 예시

승인 예시	반려 예시(오검출)
	
	반려사유 : 신호등, 표지판을 자동차로 라벨링

- 바운딩 박스는 올바르게 객체를 감싸고 있는가?

※ 객체와 박스간의 간격이 박스 너비, 높이의 2%를 초과하는 경우 반려

〈표 Ⅲ-152〉 박스정확도 반려 사유 예시

승인 예시	반려 예시(박스 정확도 부족)
	

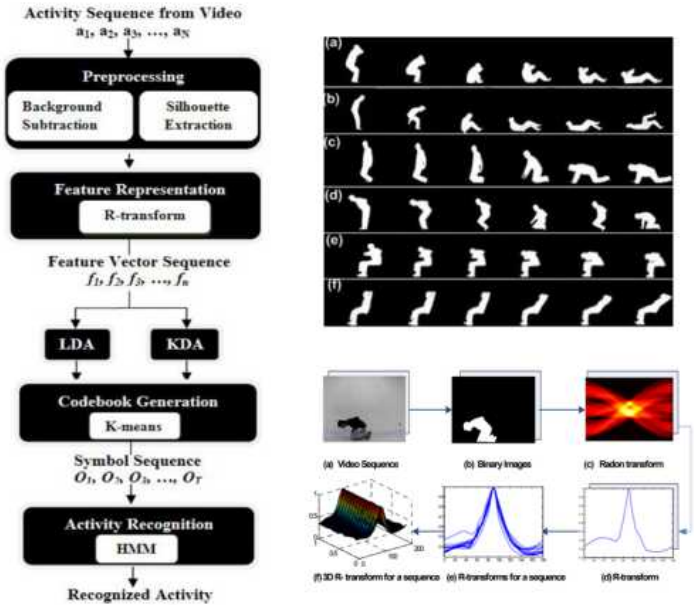
5 데이터 활용 방안

5.1 학습 모델

- 이상행동 영상데이터 인공지능 모델
  - 이상행동 인식에 대한 인공지능 모델은 다음의 4가지로 분류
    ① 실루엣 기반 이상행동 인식 모델
    ② 이동궤적기반 이상행동 인식 모델
    ③ 스파스 코딩 방식 이상행동 인식 모델
    ④ 프레임 예측 기반 이상행동 인식 모델
  - 3차원 모델 기반 이상행동 인식 알고리즘은 다중 입력프레임을 쌓아 하나의 입력 형태로 만들고 convolution 필터를 3D 큐브 형태로 만들어 시·공간 정보를 동시에 학습하는 방법으로 기존의 2D 모델을 시간축으로 팽창시켜 영상에 활용이 가능

1) 실루엣 기반 이상행동 인식 모델

- 실루엣 기반 이상행동 인식 모델은 사람과 같은 객체의 영역을 분리하여 이진 영상을 생성 후 이사행동 징후에 관련한 특징을 추출하여 HMM(Hidden Markov Model) 등의 분류기를 통해 이상행동 유무를 구별하는 방식
- 이상행동 인식 방법의 강건성은 영상에서 이상행동의 직후의 특징이 얼마나 효율적으로 추출이 되는가에 의존. 경희대학교의 연구진(Khan and Sohn)은 2011년 IEEE Consumer Electronics에서 게재한 'Abnormal human activity recognition system based on R-transform and kernel discriminant technique for elderly home care'에서 Radon Transform을 KDA(Kernel Discriminant Analysis)와 통합하여 이상행동 인식을 위한 특징추출에 활용하였으며, 인식률 향상을 위하여 첫 번째 계층에서 Radon Transform과 KDA를 이용한 이진 실루엣 추출을 수행하고, 두 번째 계층에서 k-means clustering 알고리즘과 HMM(Hiden Markov Model) 분류기를 통한 계층적 인상행동 인식 시스템을 설계 적용

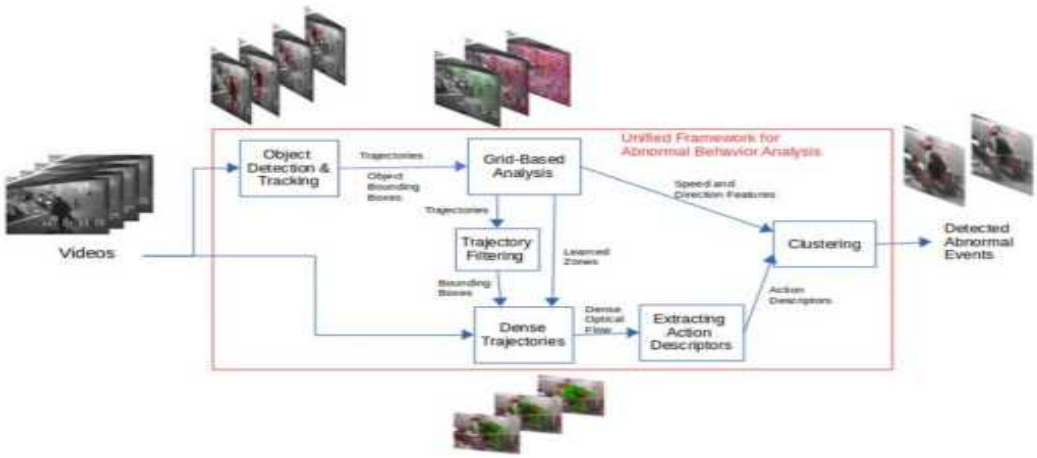


[그림 Ⅲ-169] 실루엣 기반의 이상행동 인식 기법(예시)

2) 이동궤적 기반 이상행동 인식 모델

- 고전적인 이상행동 인식 모델로 객체 추적 방법을 활용하여 사람이나 물체를 검출 및 추적하고 객체의 동선이나 자세 변화 등을 통해 이상행동을 정의하는 방법
- CCTV영상과 같이 카메라가 고정되어 있는 환경에서는 영상의 배경이 분리가 가능하며 이를 통해 객체를 검출하고 궤적을 추적. 딥러닝 기반의 객체 감지 및 추종 방법을 통한 이동궤적 활용이 가능
- 2017년 IEEE Transaction on Circuits and Systems for Video Technology에 발표된 'Toward abnormal trajectory and event detection in video surveillance'에서는 물체의 이동궤적과 픽셀기반의 접근법을 통합하여 이상행동 검출에 활용
- 객체의 이동 상태에 따라 이동행동을 감지하는 직관적 방식으로서 이동궤적으로 분명하게 정의할 수 있는 이상행동에 대한 인공지능 모델의 개발이 쉬움. 예를 들어 침입, 배회 등의 이상행동은 사람의 동선을 추적하여 진입하지 않아야하는 영역에 들어가거나 일정 영역을 배회하는 등의 동작을 검출 및 추적하여 인식하기에 용이
- 단점으로 객체의 이동궤적의 추적을 방해하는 요소가 들어간 혼잡한 장면에서 취약할 수 있음. 예를 들어 다수의 사람이 등장하여 가림이 발생하거나 객체가 섞이는 경우 추적이 어렵고 조명에 의해 그림자가 생기거나 영상이 급변하는 것에 대해서도 취약할 수 있음

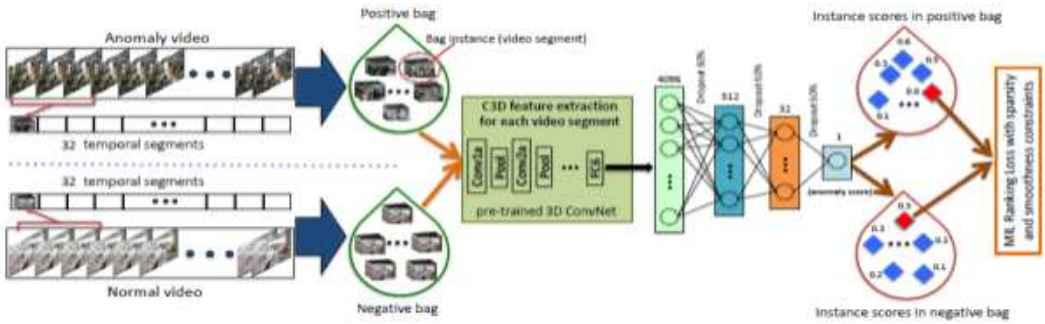
- 일반적으로 이동궤적만으로 이상행동을 정의할 수 있는 경우는 많지 않으므로 이동궤적만으로 정의하기가 모호한 이상행동을 검출하기에는 부적합



[그림 Ⅲ-170] 객체의 이동궤적 기반 이상행동 탐지기법 개요

3) 스파스 코딩 방식 이상행동 인식 모델

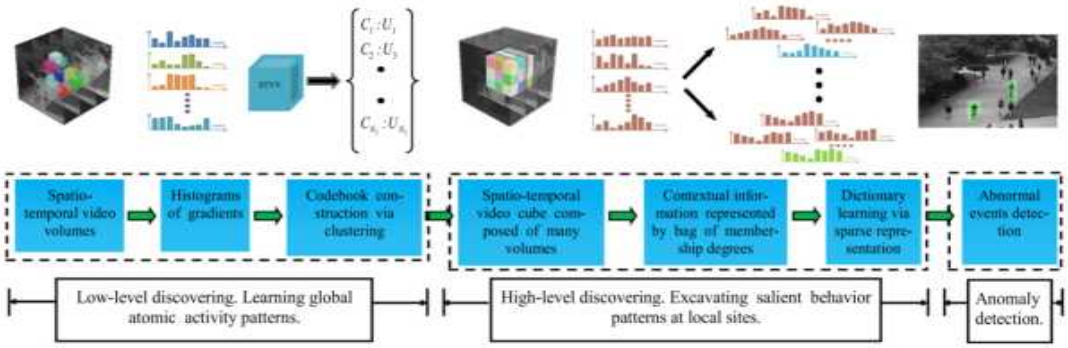
- 영상의 이상행동 구간을 검출하기 위하여 오토인코더(autoencoder)와 같은 기법을 활용하여 장면 내 등장(appearance)와 모션(motion)에 관한 특징을 추출해내고, 추출된 특징을 이용하여 이상행동을 판단하는 방식으로 기존 방식이 가지고 있던 제한점인 무제한적인 특성에서 데이터의 수집이 매우 큰 비용이 든다는 점을 해결하기 위해 고안된 인식 방법



[그림 Ⅲ-171] sparse coding 기반의 이상행동 인식 알고리즘 개념도

- 2018년 CVPR에 발표된 'Real-world anomaly detection in surveillance videos'에서는 3D Convnet을 활용하여 장면 내 모양과 모션에 대한 두가지 특징을 추출하고, 추출된 특징의 이상행동을 판단하는 네트워크를 구성

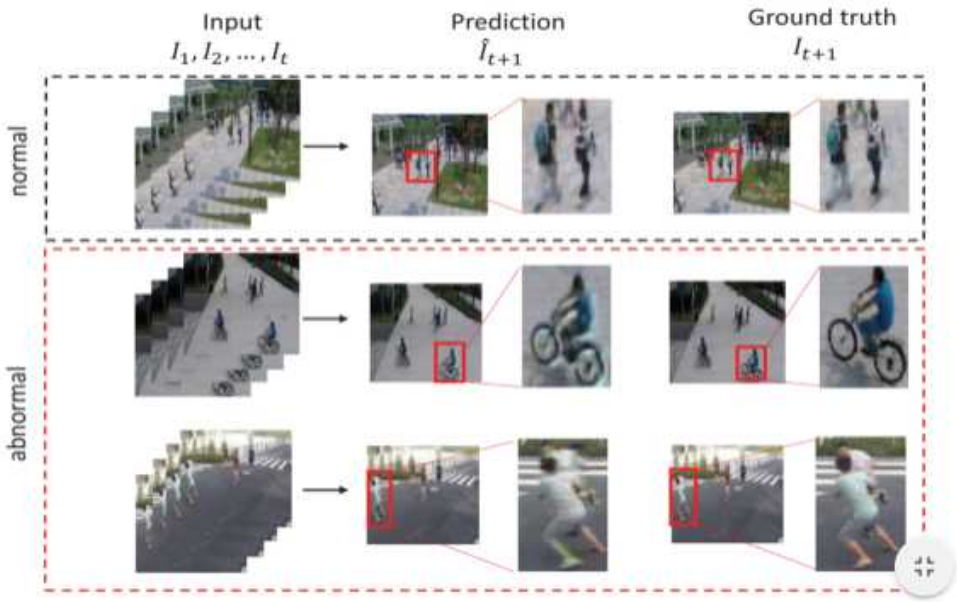
- 시공간적 영상을 활용하는 이상행동 인식 기법은 단일 영상에 의존하는 실루엣 기반의 이상행동 인식 기법에 비해 다양한 동작들을 인식하는 것에서 더 높은 성능을 나타냄. 중국 심천의 연구진은 캡처된 비디오 영상에서 이상행동 인식 및 지역화를 위해 시공간 컨텍스트의 세부 정보를 활용. 비지도통계학습(unsupervised statistical learning) 구조를 개발하여 전역 행동 패턴과 국소 행동 패턴을 검출하는 방식으로 시공간 영상 볼륨을 활용하였으며, 시공간 특징을 검출하기 위한 클러스터링 기반으로 코드북을 학습하며, 이러한 코드북으로부터 이상행동을 탐지하도록 하는 분류기를 개발
- 단점으로 비지도 학습에 가까운 약한 지도학습 방식으로 이상행동에 대한 프레임 구간을 알아서 판단하지만 일반화를 위해서는 한 가지 상황에 대해 방대한 양의 데이터를 요구



[그림 Ⅲ-172] 시공간 영상 볼륨 기반의 이상행동 탐지 기법

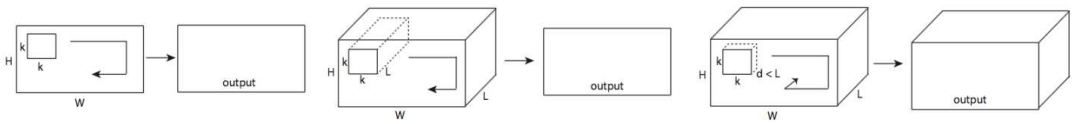
4) 프레임 예측 기반 이상행동 인식 모델

- 학습된 인공지능 모델을 기반으로 미래의 프레임을 예측하여 실제 값과 일정 차이가 발생하면 이상행동으로 감지하는 방식으로 정상 상황에 관한 데이터만을 이용하여 인공지능 모델을 학습하고 학습된 데이터로 미래 프레임을 예측하는 방식



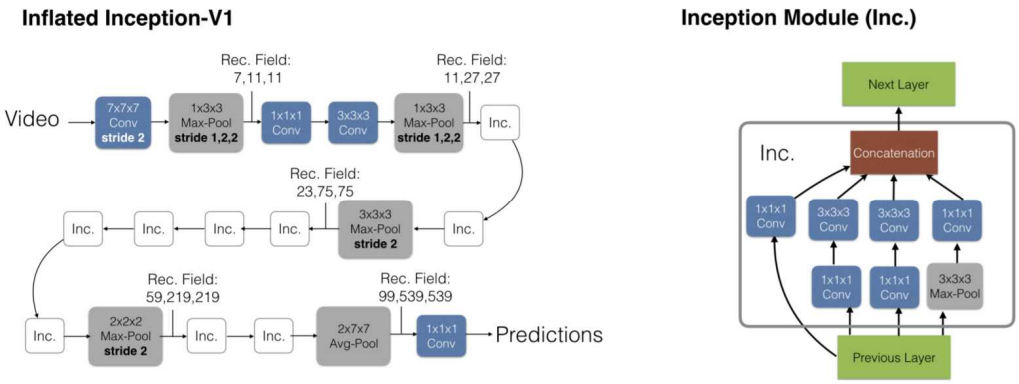
[그림 Ⅲ-173] 영상 프레임 예측 기반 이상행동 인식 예시

- 단점으로 학습된 정상 상황에 대해서만 정확한 예측이 가능하므로, 정확한 예측을 하지 못한 경우를 이상행동으로 정의하여 인식. 정확한 미래 프레임을 생성하기 위해 GAN(Generative Adversarial Network)를 사용하는 것이 특징
- 검출해야하는 이상행동이 분명하고 집중된 경우에 활용 가능하지만 정상상황에 대한 데이터가 방대하게 필요함. 따라서 학습 데이터의 양에 성능이 좌우되며, 어디까지를 이상행동으로 볼 것이냐에 대한 기준이 별도로 필요



[그림 Ⅲ-174] 2D/3D CNN 개념도

- 2017년 CVPR에 발표된 'Action Recognition? A New Model and the Kinetics Dataset' 에서는 Kinectics 데이터셋과 3D convolution 기반의 모델을 공개하였으며, 딥러닝 모델 Inception 구조를 시간축으로 확장하여 80.9%의 높은 성능을 보임



[그림 Ⅲ-175] Inflated Inception의 구조와 Inception model

- 해당 논문에서는 ImageNet에서 활용되었던 2D Inception modeld을 시간축으로 팽창시켜 그대로 비디오 영상에 적용할 수 있음을 보임

● 이상행동 학습모델 검증

- 구축한 데이터셋은 영상 클립과 라벨링 데이터로 구성되어 있으며, 이상행동 구간에 대한 라벨정보를 바탕으로 인공지능 모델을 학습하고 그 성능을 평가하여 데이터셋과 레이블의 유용성을 검증
- 이동 궤적 기반으로 이상행동 판별이 가능한 데이터셋(배회 등)의 경우 이동궤적 기반의 방식을 활용하며, 이동 궤적 만으로 판단이 어려운 이상행동 데이터의 경우 3D CNN 및 스파스 코딩방식의 모델을 활용하여 성능을 검증
- 제작한 데이터 DB의 80%를 사용하여 인공지능 모델의 학습에 활용하며, 나머지 20%의 데이터 DB를 사용하여 추론시 다음의 목표성능지표 달성 여부를 확인
- 모델 성능 지표는 KISA에서 제시한 지능형 CCTV 인증 평가 기준 활용

● 객체 추적 영상 데이터

1) 사람 탐지 및 추적, 서로 다른 카메라에서 재인식 인공지능 모델

- 사람 추적 인공지능 모델은 총 2가지 방법으로 분류
  ① 사람 탐지와 재인식을 동시에 진행하는 One-stage 방법
  ② 사람 탐지와 재인식을 순차적으로 진행하는 Two-stage 방법
- 서로 다른 카메라에서 사람 재인식 인공지능 모델을 2가지 과정으로 분류
  ① 사람 탐지 모델

② 사람 정보를 2d 좌표 공간에 Embedding 후 거리 기반 재인식 모델

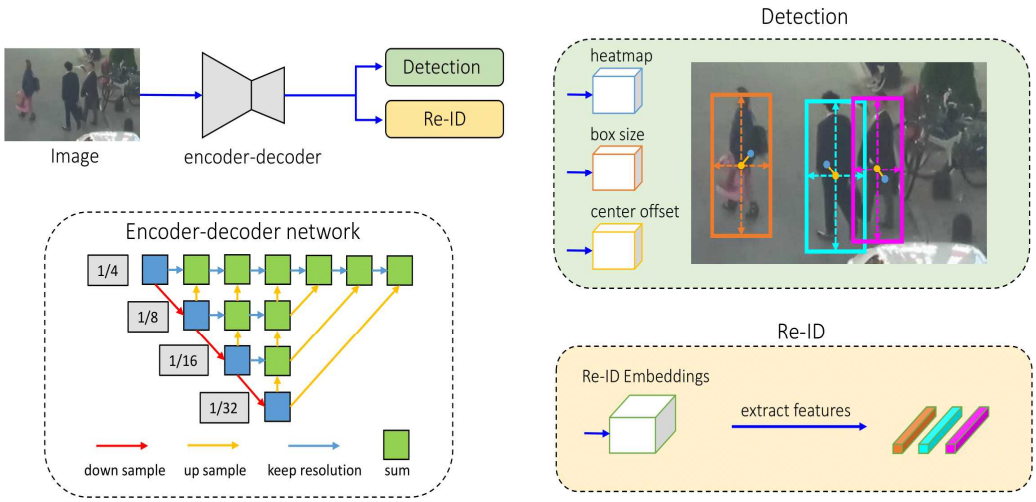
- 사람 탐지 및 추적, 서로 다른 카메라에서 재인식 인공지능 모델은 전부 2D 이미지에서 작동하는 알고리즘이지만, 가벼운 네트워크 설계로 프레임 단위로 추론하여도 CCTV에서 실시간 활용 가능
- 사람 탐지 및 추적 모델과 서로 다른 카메라에서 재인식 인공지능 모델 모두 사람을 탐지 알고리즘이 필요하다는 점에서, 서로 중복되는 네트워크를 공유하여 가벼운 네트워크 설계

2) 사람 탐지와 재인식을 동시에 진행하는 One-stage 방법

- 프레임 정보에서 다중 사람 추적을 위하여, 다중 사람 탐지 모델을 필수적이다. 따라서 사람 추적을 하기 전에, 프레임 안에 존재하는 사람의 위치 정보를 먼저 탐지해야 한다. 사람 탐지 알고리즘은 사람의 위치 정보와 분류를 동시에 하는 One-stage 방법과 사람의 위치 정보와 분류를 순차적으로 수행하는 Two-stage 방법이 존재한다. One-stage 알고리즘 중에서 가장 성능이 높다고 알려진 One-stage의 대표적인 방법으로는 2020년도에 arXiv에 공개된 `FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking‘ 이다. 그림 38에 보이는것과 같이 FairMOT는 사람의 위치 정보와, 사람의 분류(사람의 식별번호)가 하나의 네트워크에서 동시에 수행된다. 이러한 방법은 Multiple Object Tracking on MOT15,16,17,20에서 모두 State of the Art를 달성하였다. 결론적으로 One-stage 방법은 사람의 식별번호 분류를 위하여 사람의 모양, 질감등의 정보를 2차원 공간의 Embedding하는 것과 사람의 위치 정보를 탐지하는 것을 동시에 진행하는 것을 의미한다. 사람 추적에서 One-stage방법의 장점은 Two-stage방법보다 성능이 월등하게 높다는 점이다. 우리의 CCTV환경에서는 사람 추적기술은 가장 근본적이고 성능에 가장 큰 영향을 주기 때문에 우리는 성능이 높은 One-stage방법을 채택하였다

3) 사람 탐지와 재인식을 순차적으로 진행하는 Two-stage 방법

- Two-stage 사람 추적 알고리즘에서 가장 성능이 높다고 알려진 알고리즘은, 2020 European



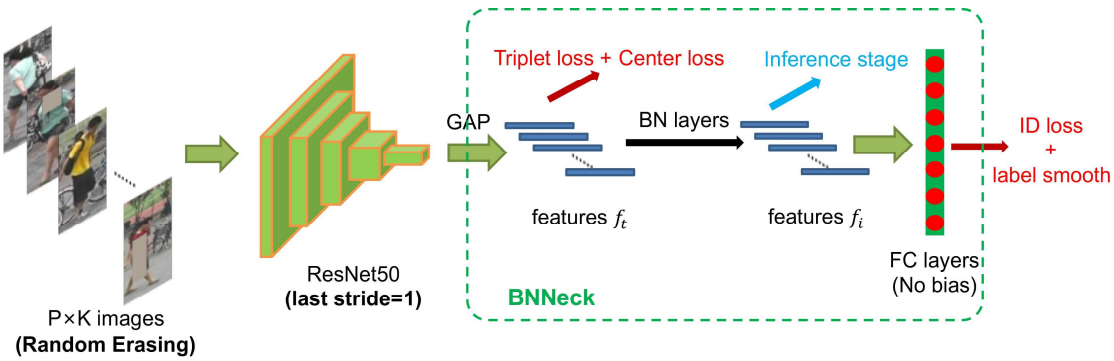
[그림 Ⅲ-176] FairMOT 모델의 알고리즘 흐름도

- Conference on Computer Vision(ECCV)에 개제된 `Towards Real-Time Multi-Object Tracking' 이다. 그림 39는 Towards Real-Time Multi-Object Tracking의 알고리즘 흐름도이다. 그림에서 보이는 것과 같이 먼저 사람의 위치 정보를 탐지하는 알고리즘이 작동하고 그 후 사람의 식별번호를 분류하기 위하여 추출한 Embedding정보를 사람의 위치 정보에 매칭시킴으로 Two-stage로 작동한다. 결론적으로 Two-stage방법은 사람의 위치를 탐지하는 것과, 식별번호를 분류하는 것이 순차적으로 진행된다. Two-stage 방법의 장점은 사람 탐지 모델이, 사람 재인식 모델에도 필요하기 때문에, 네트워크를 공유 할 수 있어 실시간의 활용이 가능함. 우선적으로 우리의 알고리즘에서는 성능이 높은 One-stage방법을 채택하였으나, 추후 네트워크 경량화등을 고려하여 Two-stage방법도 고려한다

[그림 Ⅲ-177] 사람 재인식 흐름도

4) 서로 다른 카메라에서 사람 재인식 인공지능 모델

- 서로 다른 카메라에서 사람 재인식은 우리의 CCTV영상에서 필수적이다. 그림 40에 보이는 것과 같이 사람을 위치를 탐지하여 추출된 이미지에서(Gallery) 재인식 대상(휠체어 이동자, 시각 장애인, 유모자 이동자, 주취자, 잡상인, 마취학 아동)이미지와 매칭되는 것을 찾는다. 따라서 사람의 위치를 탐지하여 추출된 이미지를 구성하는(Gallery)가 매우 중요한데, 그것은 앞에서 언급한 사람 추적결과를 재사용하여, 연산량을 줄인다. CNNs를 활용한 사람 재인식 성능은 매우 발달 되고 그 방법이 어느 정도 정형화 되었는데, 그것을 분석하고 정리한, 2020년 EEE Transactions on Multimedia에 개제된 'Bag of Tricks and A Strong ReID Baseline'은 Market1501 공개 데이터셋에서 95%의 성능을 달성하였다. 그림 41은 'Bag of Tricks and A Strong ReID Baseline'의 알고리즘 흐름도이다. 그림에 보이는 것과 같이 사람 재인식 모델에 추가적인 파라메터나, 연산량 없이, 단순 학습 기법의 변경으로 높은 성능을 달성하였다. 이러한 방법은 우리의 CCTV환경에서도 적용할 수 있는데 하나의 GPU에 여러개의 모델이 부착되어 작동하는 우리의 환경에서, 파라메터나 연산량의 증가없이 높은 성능을 낼 수 있기 때문에 적합하다

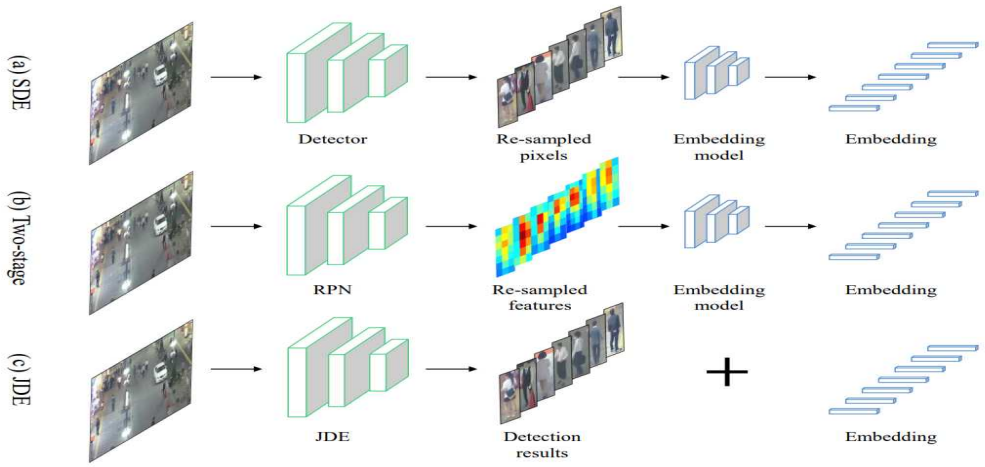


[그림 Ⅲ-178] Bag of Tricks and A Strong ReID Baseline 알고리즘의 흐름도

5) 사람 탐지 및 추적, 서로 다른 카메라에서 재인식 인공지능 학습모델 검증

- 구축한 데이터셋을 활용하여 프레임 단위로 라벨링된 정보를 바탕으로 인공지능 모델을 학습하고, 구축한 데이터 셋에서 성능을 평가하여 데이터셋과 레이블의 유용성을 검증
- 제작한 데이터 DB의 80%를 사용하여 인공지능 모델의 학습에 활용하며, 나머지 20%의 데이터 DB를 사용하여 추론시 다음의 목표성능지표 달성 여부를 확인

- 객체 추적은 multiple object tracking accuracy(MOTA)를 활용한다
- 객체 재인식은 Rank-1을 사용한다



[그림 Ⅲ-179] Towards Real-Time Multi-Object Tracking의 알고리즘 흐름도

5.2 서비스 활용 시나리오

- 도시철도 역사 이상징후 감시 및 알림서비스 테스트베드 구축
- 도시철도 역사 내 영상 획득 및 이상행동 인식 시스템 구축
- 테스트 베드 프로세스는 역사 내 CCTV 영상데이터를 전송받아 현장에 설치된 고성능 인공지능 연산용 서버를 이용해 실시간으로 이상행동 징후를 분석하고, 이상상황이 발생한 경우 송출하여 관리자에게 상황을 통지하는 방식의 서비스에 활용
- 구축데이터를 활용하여 이상행동을 탐지하고 추적데이터를 축적하여 딥러닝을 통한 사회적 약자와 서비스를 필요로 하는 대상자에게 서비스 제공
- 인공지능 관련 산업의 지속적 확대와 기술의 고도화로 다양한 영역에서 영향력 향상
- 기 구축된 시스템과 상호작용하여 수요자에게 맞춤형 서비스 제공
- 지차체 중심으로 성과확산사업을 연속적으로 실시할 수 있는 기회 제공
- 참여기업은 중장기적으로 사업의 성과를 전국적으로 확산하기 위한 사업화를 계획하고 추진
- 수행기관(대전도시철도공사) 역사에 설치하여 지속적인 유지관리 추진
- 과제를 위해 설치한 고성능 CCTV로 영상을 받고 응용프로그램으로 딥러닝하여 서비스를 제공하며 이용객 등의 의견을 참조하여 지속적인 업데이트 진행

제15장 패션상품 및 착용 영상 AI 데이터

1 데이터 정보 요약

1.1 가이드 분류

대분류	비디오	중분류	비전	소분류	MP4

1.2 데이터 정보

데이터 이름	패션상품 및 착용 영상 AI 데이터
활용 분야	• 패션상품 검색 • 패션상품 세부 영역 추정 • 패션상품 신규 디자인 생성 • 패션상품 모델 착용 영상생성 • 패션모델 자세 추정 • 패션모델 세부 영역 추정
데이터 요약	본 데이터는 패션상품과 그들 중 일부를 조합해서 착용한 실제 착용 영상의 쌍(pair)를 수집하고, 각 데이터에서 의미론적 영역 및 특징점을 추출하여 구축함

1.3 데이터 구축 개요

[그림 Ⅲ-180] 학습용 데이터 구축 공정도

1.4 구축 목적

- 한국 오프라인 패션시장을 주도했던 동대문의 많은 소상공인들은 공간적 제약이 없어 글로벌 시장까지 바라볼 수 있는 이커머스 시장으로 활발하게 진출하고 있고, 최근 COVID-19의 여파로 그 현상은 더욱 가속화되고 있음
- 최근 소셜미디어의 발달로 인플루언서의 패션상품 제작 및 판매 시장이 형성, 점차 인플루언서 자체가 브랜드화 되고 있음
- 인플루언서 본인이 스스로 촬영 모델이 될 수 있는 경우 비용적인 측면에서 우위를 점할 수 있지만, 대부분의 소상공인들 및 인플루언서들은 비용적인 문제로 패션 스튜디오 촬영을 수행하지 못하고 있는 실정임

1.5 활용 분야

- 판매 목적으로 수급한 패션 아이템의 사진을 활용하여 실제 해당 아이템을 착용한 착용 영상을 생성함으로써 상품 홍보 영상을 인공지능 기술을 활용하여 제작
- 패션 상품 영상을 조합하고 이를 활용하여 패션 착용 영상을 생성함으로써 여러 패션상품 조합을 반영한 패션 코디 시뮬레이션 가능

1.6 유의 사항

- 사진에 대한 저작권을 가진 사진작가, 초상권을 가진 모델, 디자인에 대한 권리를 가지고 있는 옷 제작자들을 모집하고 법적인 문제를 자문받아 해결하여 사진 형태의 데이터 수집 및 공개 시 발생 가능한 위험 요소를 해결한 데이터를 수집하여 어노테이션 작업에 사용함
- 저작권, 초상권, 공표권, 공중송신권, 동일성 유지권, 저작인격권 등 다양한 법적 문제 해결

2 데이터 획득 및 정제

2.1 원시 데이터 선정

〈표 Ⅲ-153〉 원시 데이터 정보

주요 내용	데이터 수집 방법	데이터 구축량	데이터 형식
스튜디오 패션 영상	크라우드 소싱	600만장	영상
패션제품 대표 사진	크라우드 소싱	4만장	영상
모델의 자세 데이터	크라우드 소싱	12만개	데이터 파일
모델의 semantic영역	크라우드 소싱	12만개	데이터 파일 또는 이미지
제품의 자세 데이터	크라우드 소싱	4만개	데이터 파일
제품의 semantic 영역	크라우드 소싱	4만개	데이터 파일 또는 이미지

2.2 규제관련 사항

- 관련 내용 없음

※ 본 데이터 가이드라인 상에는 규제 관련 내용이 없지만, 데이터 구축 시 규제 관련 내용 확인 및 검토 필요

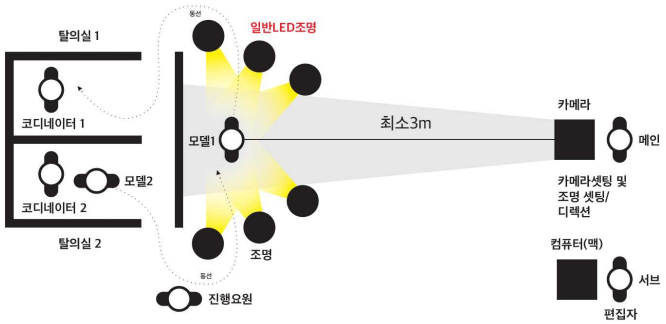
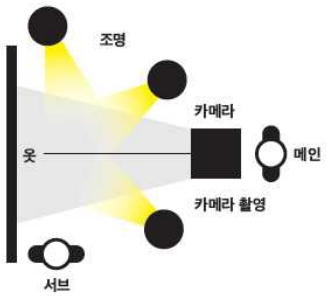
2.3 획득 및 정제 절차

〈표 Ⅲ-154〉 획득 및 정제 작업 내용

구분	내용
데이터 획득	• (패션 영상 데이터를 위한 촬영) 패션상품 사진 및 스튜디오 패션 영상 촬영시 촬영 세팅(카메라 위치, 조명 등)을 고정함으로써 정형화된 데이터가 취득되도록 하여 정제과정을 최소화 함 • 착용샷 부스별 일일 촬영 가능 수는 5분에 1벌 촬영 X 7시간 = 84벌 촬영 • 착용샷 부스는 4대를 운영하여 하루 336벌 착용샷 촬영 • 과업기간동안 90 영업일을 운영하여 총 30,000벌 착용샷 촬영 • 바닥샷 부스별 일일 촬영 가능 수는 5분에 1벌 촬영 X 7시간 = 84벌 촬영 • 바닥샷 부스는 3대를 운영하여 하루 252벌 바닥샷 촬영 • 과업기간동안 90 영업일을 운영하여 총 20,000벌 바닥샷 촬영
데이터 정제	• 원시데이터 정체 절차에는 후보정 작업이 포함되는 절차들이 포함되며 영상 전문가가 영상 툴(예.파이널컷프로)을 활용하여 정제작업을 진행함 • 획득된 영상 데이터에 대해서 동일한 영상적 특징을 가지도록 레벨 조정 등 후보정 작업을 진행함 • 영상은 특정위치를 크롭하여 활용함으로써 데이터 화 함 • 스튜디오 패션 사진의 경우 모델의 움직임을 포함하고 있으므로 흐림현상(blur)가 없는 데이터를 취득하기 위해 60fps이상의 속도로 촬영된 영상에서 각 포즈 전환당 45장의 영상을 추출하여 활용함

2.4 획득 및 정제 기준

〈표 Ⅲ-155〉 획득 및 정제 기준

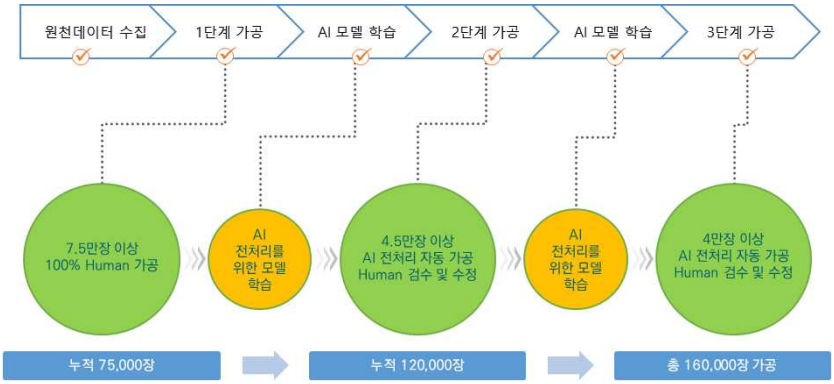
구분	스튜디오 패션 영상(착용샷) 촬영을 위한 스튜디오 세팅
획득 기준	
	• 착용샷 촬영을 위해 카메라와 피사체 간 거리가 최소 3M 이상 확보되는 공간 필요 • 동영상을 스틸컷으로 Blur 없이 저장할 수 있는 충분한 광량이 필요 • 탈의실 내 코디네이터와 모델이 2인 1조씩 2개 조로 운영하여 옷을 갈아입고 촬영하는 시간의 갭을 최소화함 • 카메라 세팅과 촬영을 진행할 전문 인력과 촬영된 데이터를 편집할 IT 인력이 필요함
	패션상품 대표 사진(바닥샷) 촬영을 위한 스튜디오 세팅
	
	• 바닥샷의 촬영을 위해 옷이 구김 없이 잘 펴지도록 하고 촬영 후 옷을 교체하는 인력이 필요하고 충분한 광량에 스튜디오 셋팅이 필요하고 촬영된 데이터를 편집할 IT 인력이 필요함 • 모델이 옷을 착용하고 스튜디오에 입장하는 순간부터 퇴장하는 순간까지 동영상으로 촬영 • 한 개의 코디 기준 5분 동안 촬영을 진행함 • 진행 시간 동안 모델이 사전에 제시된 포즈 중 4종을 골라 전환함으로써 다양한 각도에서 촬영하는 것과 동일한 효과를 얻을 수 있음 • 바닥샷은 의류 아이템의 앞, 뒷면을 기본으로 촬영을 진행함 • 모자/신발의 경우 착용 시 보이는 부분과 보이지 않는 부분으로 구분하여 촬영

구분	스튜디오 패션 영상(착용샷) 촬영을 위한 스튜디오 세팅
정제 기준	• (데이터 촬영포맷) 데이터의 촬영은 초당 120fps(frame per second)와 셔터스피드 1/500이상 조리개 8.0이상 영상 화각 50mm 이상의 렌즈를 사용하여 MOV파일 형태의 동영상으로 데이터를 저장하며 착용샷 스튜디오 하나당 하루 8시간 기준 소비되는 메모리는 1TB 정도로 예상하고 바닥샷은 동일 조건에 RAW 혹은 JPG 이미지로 촬영함 • (데이터 편집) 착용샷을 촬영한 동영상은 포즈별로 구분하여 파일을 생성 관리하고 촬영포맷을 스포츠 경기 촬영에 준하는 1/500 셔터스피드로 생성하여 동영상을 스틸컷으로 변환해도 흐림현상(blur) 없이 깨끗한 영상을 확보할 수 있고 바닥샷은 RAW 파일을 기본으로 하여 필요시 보정작업을 진행함

3 어노테이션/라벨링

3.1 어노테이션 / 라벨링 절차

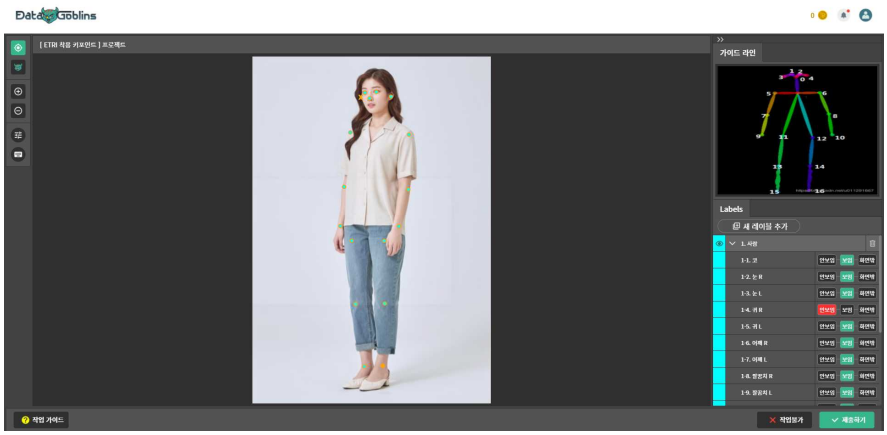
- 데이터 가공을 위해 데이터를 1차로 취득 및 휴먼 어노테이션 작업
- 해당 데이터로 주관기관이 Keypoint 및 Segmentation 전처리를 위한 인공지능 모델 구축
- 구축된 모델은 이후 작업 시 사전 결과물로 활용, 데이터 가공 기간을 단축
  - 1단계 가공 : 7.5만장 이상의 원천 데이터를 크라우드 소싱 인력을 통해 100% 가공. 인공지능 전처리를 위한 알고리즘 학습을 위한 기본 데이터세트로 활용
  - 2단계 가공 : 학습된 인공지능 전처리 모델을 적용하여 12만장 이상의 원천 데이터를 자동 레이블링 진행. 크라우드 작업자들을 통해 수정 및 검수 진행. 인공지능 전처리 모델을 정교화하기 위해 작업 결과를 추가 학습 진행
  - 3단계 가공 : 보다 정교화된 인공지능 전처리 모델을 적용하여 나머지 원천 데이터를 자동 레이블링 진행. 인하우스 작업자들을 통해 수정 및 검수 진행



[그림 Ⅲ-181] 데이터 가공 절차

3.2 어노테이션 / 라벨링 기준

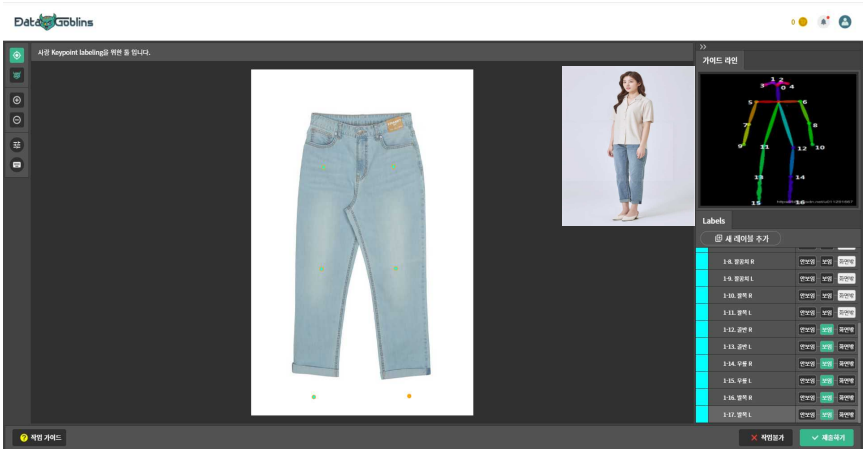
- 모델 포즈 정보(착용샷 Keypoint)
  - 포즈 정보 사진에 알맞도록 개별 포인트들의 속성을 지정함
  - 새 레이블 추가 버튼으로 새로운 키포인트 세트 생성
  - 코를 제외한 좌우가 있는 포인트의 경우 실제 포인트 위치에 맞추되, 보이지 않는 포인트(가려졌으나 추정 가능한 포인트)는 '안보임', 아예 화면 밖에 있는 포인트는 '화면밖'으로 처리



[그림 Ⅲ-182] 키포인트를 모두 추출한 모습

- 상품 포즈 정보(바닥샷 Keypoint)
  - 상품 정보 사진에 알맞도록 개별 포인트들의 속성을 지정함
  - 새 레이블 추가 버튼으로 새로운 키포인트 세트 생성

- 착용샷을 참고하여 골반, 무릎, 발목 등 관절 위치에 키포인트 생성



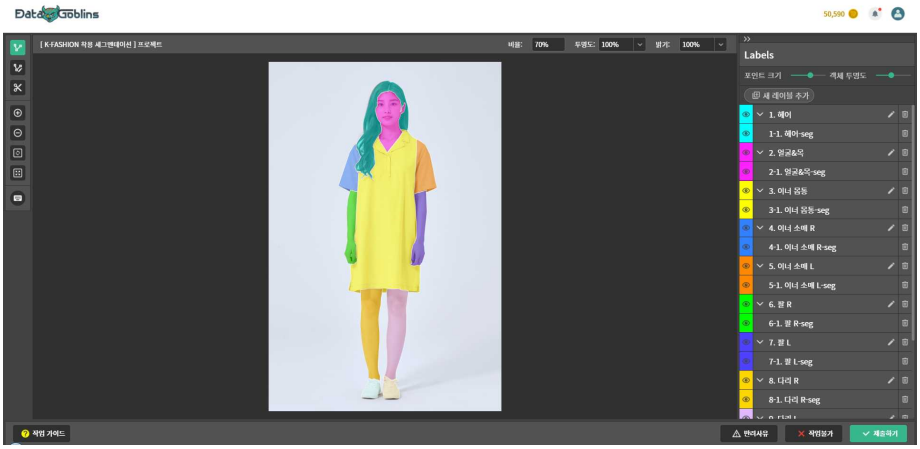
[그림 Ⅲ-183] 착용샷을 고려한 키포인트 생성

- 상품 의미론적 영역(바닥샷 Segmentation)
  - 먼저 상품 영역 전체에 Segmentation 작업
  - 가위 툴을 사용하여 영역을 분할
  - 상의 의류의 경우 오른쪽 소매, 왼쪽 소매, 안 보이는 영역 순으로 분할
  - 단계마다 분할된 영역을 적합한 레이블로 변경
  - 안 보이는 영역까지 모두 지정하여 바닥샷 Segmentation 작업 마무리

[그림 Ⅲ-184] 상품 의미론적 영역 Segmentation 작업

- 모델 의미론적 영역(착용샷 Segmentation)
  - 사전에 정의한 레이블링 순서에 따라 착용샷에 Segmentation 작업

- 사진에 따라 1개 영역을 여러 조각으로 레이블링할 수 있음
- 툴 사용 시 동일 영역이 다른 색상의 조각으로 작업되지 않도록 유의
- Segmentation 작업을 완료한 모습



[그림 Ⅲ-185] 모델 의미론적 영역 Segmentation 작업

3.3 어노테이션 / 라벨링 교육

- 작업자 조직
  - 크라우드 작업자는 100% 수동 작업이 필요한 원천 데이터 및 인공지능 전처리를 통해 1차 가공한 데이터를 수정 작업하는 단계에 투입하여 작업 품질과 양에 따라 리워드 지급
- 라벨링 조직
  - 1차 검수자(크라우드 검수자)는 데이터 통과 여부를 판단하여 통과 혹은 반려하며, 크라우드 작업자에 의해 작업된 데이터 반려 및 재검수 작업
  - 2차 검수자(프로젝트 관리자 인력 포함)는 검수 및 오작업에 대한 수정 작업
  - 프로젝트 관리자는 데이터의 전반적인 품질을 관리하며, 모든 데이터에 대한 최종 검수 작업을 진행
- 필수 관련 교육
  - 어노테이션을 수행하는 전문의 및 외부 수행자를 대상으로 어노테이션 툴(저작 도구) 사용 방법, 어노테이션 방법 등 교육 진행

〈표 Ⅲ-156〉 필수 관련 교육 내용

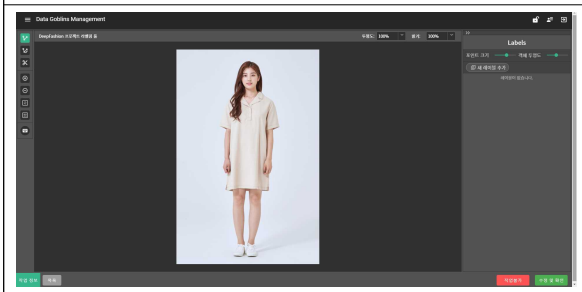
교육 구분	내용	수행 기간
인공지능과 데이터	• 인공지능에서 데이터의 중요성 이해 • 데이터 가공의 이해	1시간
데이터 가공 실습	• Data Goblins 플랫폼의 소개와 사용법 • Data Annotation 실습 (1) 이미지 작업 도구 기본 기능 실습 • Data Annotation 실습 (2) Keypoint 작업 실습 • Data Annotation 실습 (3) Segmentation 작업 실습	3시간
데이터 검수 실습	• Data Annotation 품질관리의 이해 • Data Annotation 품질관리 프로세스	1시간

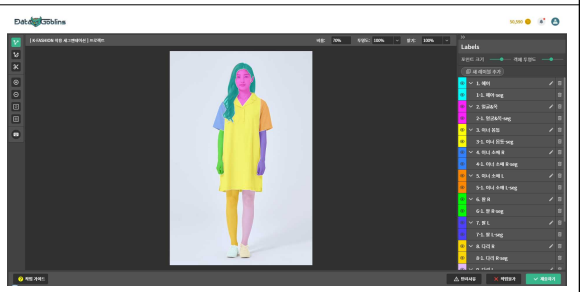
3.4 어노테이션 / 라벨링 도구 및 사용법

- Keypoint 및 Segmentation 작업 도구

Keypoint 작업 도구

- 새 레이블 추가 버튼으로 새로운 Keypoint Set 추가
- 적합한 위치에 포인트 작업, 포인트 개별 속성 부여 가능





Segmentation 작업 도구

- 새 레이블 추가 버튼으로 새로운 Segmentation Set 추가
- 전체 영역 작업, 가위 툴로 영역별 분할 작업 가능
- 분할한 영역 적합한 부분으로 이동 가능

[그림 Ⅲ-186] Keypoint / Segmentation 작업 도구

4 데이터 검수

4.1 검수 절차

● 기본 검수 절차

1) 검수자 검수

- 가공된 작업물을 임의의 작업자에게 할당하는 방식의 검수를 통해 데이터의 오류 검사
- 가공된 데이터의 품질을 관리하는 품질 관리자를 지정 배치, 재교육 및 가이드 문서 수정

2) 전문가 검수

- 가공된 데이터를 최종 확인 및 수정하는 전문 검수자 지정 배치, 데이터 오류율 최소화
- 작업 품질이 우수하거나 검수자 교육에서 우수한 성적을 나타낸 경우 전문 검수자로 선정

3) 검수 요청/재작업

- 검수 기준에 맞지 않는 데이터를 사유와 함께 반려, 작업자 재작업 및 검수
- 지속적인 구축 가이드라인 효율화 및 수정에 반려 데이터 활용
- 최종적으로 TTA와의 협업을 통해 데이터 검증절차를 진행

● 최종 검수 및 승인

1) 작업 결과 제공: 주관기관에서 정의한 최종 출력 포맷에 맞도록 데이터 가공 및 제공

2) 결과 승인: ETRI에서 최종 데이터 검수 및 승인

3) 프로젝트 종료

4.2 검수 기준

〈표 Ⅲ-157〉 검수 기준

번호	어노테이션 내용	재작업 항목
1	모델 포즈 정보	• Keypoint 속성이 잘못 지정된 경우 • Point 위치가 바르지 않은 경우
2	상품 포즈 정보	• Keypoint 속성이 잘못 지정된 경우 • Point 위치가 바르지 않은 경우
3	모델 의미론적 영역	• Segmentation 영역이 올바르지 않은 경우 • Segmentation 영역 개수가 틀린 경우
4	상품 의미론적 영역	• Segmentation 영역이 올바르지 않은 경우 • Segmentation 영역 개수가 틀린 경우

5 데이터 활용 방안

5.1 학습 모델

- 스튜디오 패션 영상 생성 모델 학습
  - (교차 검증을 위한 학습데이터 분리) 총 데이터 중 학습 : 검증 : 평가 데이터를 6 : 1 : 3 으로 분배하여 학습 및 평가에 활용함
  - (인공지능 모델의 입력) 상품 사진 10장(대표상품 5종에 대한 2개 시점에서의 사진), 모델 자세에 대한 히트맵 (또는 영역 정보)
  - (인공지능 모델의 출력) 패션 스튜디오 영상
  - (인공지능 모델) U-Net 기반 영상생성 모델

5.2 서비스 활용 시나리오

- (상품 홍보 영상 제작) 판매 목적으로 수급한 패션 아이템의 사진을 활용하여 실제 해당 아이템을 착용한 착용 영상을 생성함으로써 상품 홍보 영상을 인공지능 기술을 활용하여 제작
- (패션 코디 시뮬레이션) 패션 상품 영상을 조합하고 이를 활용하여 패션 착용 영상을 생성함으로써 여러 패션상품 조합을 반영한 패션 코디 시뮬레이션 가능

Ⅳ
부 록

제1장 용어 정의

〈 ㄱ 〉

- 검증 데이터(셋) (Validation Data Set)
  - 전체 학습데이터셋 중에서 일정한 비율을 정하여 인공지능의 기계학습에 따른 성능을 보정하거나 향상하는 용도로 사용하는 데이터셋
- 광학문자인식 (OCR, Optical Character Recognition)
  - 사람이 쓰거나 기계로 인쇄한 문자의 영상을 기계가 읽을 수 있는 문자로 변환하는 것
- 기계학습 (Machine Learning)
  - 인간이 자연적으로 수행하는 학습 능력과 같은 기능을 컴퓨터에서 실현하려는 기술이나 방법

〈 ㄷ 〉

- 데이터 획득 (Data Acquisition)
  - 인공지능의 기계학습에 필요한 데이터를 현실 세계에서 직접 수집 또는 생성하거나, 이미 보유하고 있는 조직이나 시스템 등으로부터 법률적 제약이 없도록 '원시데이터'를 확보하는 활동
- 데이터 정제 (Data Refinement)
  - 획득한 원시데이터를 기계학습에 필요한 형식으로 맞추거나 불필요한 중복을 제거하며, 개인정보를 비식별화하여 처리하는 등 일련의 전처리 과정을 통해 '원천데이터'를 확보하는 활동
- 데이터 라벨링 (Data Labeling)
  - 인공지능이 기계학습에 활용할 수 있도록 기능이나 목적에 부합하는 정보를 원천데이터에 부착하는 활동

- 데이터 학습 (Data Machine Learning)
  - 학습데이터셋의 훈련데이터셋, 검증데이터셋을 이용하여 선정된 인공지능 알고리즘을 학습시키고, 학습된 인공지능 모델의 성능을 향상시키거나 보정하는 활동

〈 ㄹ 〉

- 라벨링데이터 (Labeled Data)
  - 원천데이터에 부여한 '참값', 파일형식이나 해상도 등의 속성, 그리고 설명이나 주석 등이 포함된 '어노테이션'의 집합

〈 ㅇ 〉

- 어노테이션 (Annotation)
  - 데이터 라벨링 시 원천데이터에 주석을 표시하는 작업을 의미하며, 추가 부착되는 설명정보 데이터는 기능 목적에 따라 다양한 형태로 표현될 수 있으며 이러한 설명정보 표현방식을 지칭

    ※ 용어사용 예 : 사물 바운딩박스 어노테이션, 클래스 라벨링 어노테이션 등
- 원시데이터 (Raw Data)
  - 기계학습을 목적으로 획득 단계에서 수집 또는 생성한 텍스트, 이미지, 비디오, 오디오 등의 데이터
- 원천데이터 (Source Data, Unlabeled Data)
  - 원시데이터를 라벨링 공정에 투입하기 위해 필요한 전처리 등 정제 작업을 수행한 데이터로 라벨링데이터가 부여되지 않은 상태의 데이터
- 인공지능 (Artificial Intelligence)
  - 자연 언어의 이해, 음성 번역, 문제 해결, 학습과 지식 획득, 인지 과학 등에 응용하기 위해 인간의 지능이 갖는 학습, 추리, 적응, 논증 등의 기능을 갖춘 컴퓨터 시스템
- 인공지능 학습용 데이터 구축
  - 임무정의, 데이터 획득, 데이터 정제, 데이터 라벨링 등 인공지능 학습용 데이터를 구축하는 일련의 활동
- 임무 정의 (Task Define)

- 인공지능 기계학습을 통해 해결하고자 하는 문제를 명확하게 정의하고, 문제 해결에 필요한 학습용 데이터의 요구사항을 구체적으로 정의하고 설계하는 활동

〈 ㅊ 〉

● 참값 (Ground Truth)

- 인공지능의 기계학습 목적에 따라 원천데이터에 라벨링된 정확한 값이나 사실의 의미적 표현

〈 ㅋ 〉

● 크라우드소싱 (Crowdsourcing)

- 대중(crowd)과 아웃소싱(outsourcing)의 합성어로 기업 활동의 일부 과정에서 일반 대중(크라우드워커)을 참여시키는 것을 의미

● 크라우드워커 (Crowd worker)

- 일반인이 기업의 업무 용역을 대행 수행하고, 일정 대가를 받는 경우를 의미하며, 집이나 재택근무 등의 형태로도 업무 수행이 가능하고, 자유롭게 할당된 과제물을 수행하는 일자리

  예) A 참여기관이 '음식데이터 AI구축과제'를 추진하는 경우 '크라우드워커'를 고용하여 일반인으로부터의 음식 사진 데이터를 대량 수집

〈 ㅎ 〉

● 학습데이터(셋) (AI Data Set)

- 인공지능의 기계학습에 사용하는 원천데이터와 라벨링데이터의 묶음을 말하며, 사용하는 목적에 따라 '훈련데이터셋', '검증데이터셋', '시험데이터셋'으로 구분

● 훈련데이터(셋) (Training Data Set)

- 전체 학습데이터셋 중에서 일정한 비율을 정하여 인공지능의 기계학습에 직접 사용하는 데이터셋

● 학습모델 임무 (Learning Model TASK)

- 학습모델 알고리즘의 목적을 구분하기 위한 단위로서, 지능별 기술*에 따라 '시각지능', '언어·청각지능', '음성지능', '융합지능' 등 지능별 기술에 따라 다양하게 분류할 수 있음

* 출처: NIA, AI INSIGHT REPORT('19.12.)

- 다만, 본 『품질관리 안내서』에서는 지능정보원에서 공모하는 '인공지능 학습용 데이터 구축사업'의 구축사례들을 기반으로 널리 활용되는 Task 분야를 정리하여 '언어(이해/생성)지능' 및 '시각(인지/생성)지능'으로 분류함

제2장 구축계획서 작성요령

- ‘제2장. 구축계획서 작성 요령’에서는 과기부와 지능정보원에서 공모하는 ‘인공지능 학습용 데이터 구축사업’을 수행하는데 필요한 ‘구축계획서’의 작성을 위해, 제안사가 참고할 수 있도록 계획서 양식 및 작성 요령을 안내한다. 특히, ‘1-Cycle 진행방안’과 ‘조직 및 인력 관리방안’ 관련 구축계획서 작성을 통해, 사업자가 불필요한 시행착오를 줄일 수 있도록 안내한다.

1 구축 개요

1.1 구축 배경

- 작성 양식 (예시)

〈표 Ⅳ-1〉 구축 배경

구분	내용
구축 배경	• 인공지능 학습용 데이터 구축 배경 작성
구축 필요성	• 인공지능 학습용 데이터 구축 필요성 작성

〈 작성요령 〉

• 인공지능 학습용 데이터 구축 배경, 필요성 등을 제시한다.
• 인공지능 학습용 데이터 구축을 위한 전체적인 설명을 요약하여 작성한다.
• [참고] 『제1권. 품질관리 안내서』 내, ‘Ⅰ-제1장’

1.2 구축 목적

● 작성 양식 (예시)

〈표 Ⅳ-2〉 구축 목적

구분	내용
구축 목적	• 인공지능 학습용 데이터 구축 목적 작성
활용 방안	• 인공지능 학습용 데이터 구축에 따른 활용 방안

〈 작성요령 〉

• 인공지능 학습용 데이터 구축 목적 및 구축에 따른 활용 방안 등을 제시한다.
• 인공지능 학습용 데이터 구축 사업과 빅데이터 사업과의 차별점을 고려하여 목적을 작성한다.
• [참고] 『제1권. 품질관리 안내서』 내, 'Ⅰ-제1장', 'Ⅰ-제3장'

2 구축 데이터 정의

2.1 해당 데이터 영역 배경지식

● 작성 양식 (예시)

〈표 Ⅳ-3〉 데이터 영역 배경지식

<table>
<tr><th>구분</th><th colspan="9">내용</th></tr>
<tr><td rowspan="3">해당 데이터 영역</td><td colspan="9">• 아래 8개의 데이터 영역 중 하나를 선택하고, 속하는 영역가 없을 시 '기타' 표시</td></tr>
<tr><td>①</td><td>②</td><td>③</td><td>④</td><td>⑤</td><td>⑥</td><td>⑦</td><td>⑧</td><td>⑨</td></tr>
<tr><td>한국어</td><td>영상
이미지</td><td>헬스
케어</td><td>교통·
물류</td><td>재난·
안전·
환경</td><td>농·축
·수산</td><td>제조·
로보
틱스</td><td>문화·
스포츠
·관광</td><td>기타</td></tr>
<tr><td></td><td>☐</td><td>☐</td><td>☐</td><td>☐</td><td>☐</td><td>☐</td><td>☐</td><td>☐</td><td>☐</td></tr>
<tr><td>해당 산업분야
관련 수행 사례</td><td colspan="9">• 해당되는 데이터 영역과 관련된 사업 수행 사례를 작성하고, 시사점 도출</td></tr>
<tr><td>국·내외 동향 분석</td><td colspan="9">• 해당되는 데이터 영역과 관련된 인공지능 생태계를 분석하고, 시사점 도출</td></tr>
</table>

〈 작성요령 〉

• 제안사의 사업 수행 사례 등을 통해 해당 데이터 영역에 대한 배경지식을 작성한다.
• 국·내외 동향 분석을 통해, 산업분야 관점에서 인공지능 데이터 구축의 중요성을 작성한다.
• [참고] 『제1권. 품질관리 안내서』 내, 'Ⅴ-제2장-1절'

2.2 원시데이터 유형

2.2.1 원시데이터 정의

● 작성 양식 (예시)

〈표 Ⅳ-4〉 원시데이터 정의

구분	내용				
수집 담당 기업	• 수집 담당 기업명 작성				
원시데이터 유형	• 아래 4개의 데이터 유형 중 해당되는 유형을 모두 선택하고, 예시에 작성되지 않은 데이터 유형의 경우 추가 기재				
	① 텍스트	② 이미지	③ 비디오	④ 오디오	⑤ 기타
	□	□	□	□	□
원시데이터 포맷	• 원시데이터 파일 형식 작성				
원시데이터 규모	• 원시데이터 포맷 별 구축 규모 작성				

〈 작성요령 〉

• 본 사업에서 정의하는 원시데이터의 의미를 파악하고, RFP의 기준을 만족하도록 목표를 정의하여 작성한다.
• 여러 포맷으로 원시데이터를 획득하는 경우, 각각의 포맷 별 구축 규모를 상세히 작성한다.
• [참고] 『제1권. 품질관리 안내서』 내, 'Ⅲ-제2장', 'Ⅴ-제2장-2절'

2.2.2 수집방안

● 작성 양식 (예시)

〈표 Ⅳ-5〉 데이터 수집방안

구분	내용
획득 프로세스	• 원시데이터를 획득하는 과정 작성
데이터 획득 장소	• 원시데이터 획득을 위한 장소 및 환경 작성
데이터 획득 도구	• 원시데이터 획득을 위한 도구 및 획득 방안 작성
데이터 저장 방안	• 획득한 원시데이터를 저장하는 방법 작성
데이터 관리 방안	• 획득한 원시데이터를 대상으로 품질 관리하는 방안 작성 • 원시데이터 중 기준에 만족하지 못하는 경우 삭제하거나 재 수집하는 방안 작성

〈 작성요령 〉

• 본 사업에서 정의하는 원시데이터의 의미를 파악하고, RFP의 기준을 만족하도록 '획득', '저장', '관리'하는 방안을 작성한다.
• 원시 데이터 항목별 데이터 획득 방법, 법적문제 발생가능여부 등을 검토하여 실제로 인공지능 학습용 데이터 구축에 활용할 수 있는 데이터를 선정한다.
• 데이터 품질, 획득 가능성(가능여부 및 획득량), 획득 비용 및 기술수준, 법적 요건 등을 검토하여 획득할 데이터를 최종 선정한다.

〈 작성요령 〉

- 작업자에게 제공할 '수집 가이드라인'에 대한 작성 방안을 기재한다. 해당 가이드라인은 선정된 원시 데이터를 획득하기 위해 필요한 정보 또는 원시 데이터 획득현황을 파악하는데 활용한다.
- [참고] 『제1권. 품질관리 안내서』 내, 'Ⅲ-제2장'

2.3 원천데이터 유형

2.3.1 원천데이터 정의

● 작성 양식 (예시)

〈표 Ⅳ-6〉 원천데이터 정의

구분	내용				
정제 담당 기업	• 정제 담당 기업명 작성				
원천데이터 유형	• 아래 4개의 데이터 유형 중 해당되는 유형을 모두 선택하고, 예시에 작성되지 않은 데이터 유형의 경우 추가 기재				
	① 텍스트	② 이미지	③ 비디오	④ 오디오	⑤ 기타
	□	□	□	□	□
원천데이터 포맷	• 원천데이터 파일 형식 작성				
원천데이터 규모	• 원천데이터 포맷 별 구축 규모 작성				
작업 종류	• 아래 정제 작업 종류 중 해당되는 내용을 모두 선택하고, 예시에 작성되지 않은 정제 작업의 경우 추가 기재				
	① 중복제거	② 데이터 자르기	③ 비식별화	④ OCR인식	⑤ 기타
	□	□	□	□	□

〈 작성요령 〉

- 본 사업에서 정의하는 원천데이터의 의미를 파악하고, RFP의 기준을 만족하도록 목표를 정의하여 작성한다.
- 여러 포맷으로 원시데이터가 원천데이터로 정제되는 경우, 각각의 포맷 별 정제 규모를 상세히 작성한다.
- [참고] 『제1권. 품질관리 안내서』 내, 'Ⅲ-제2장', 'Ⅴ-제2장-2절'

2.3.2 정제방안

● 작성 양식 (예시)

〈표 Ⅳ-7〉 데이터 정제방안

구분	내용	
정제 프로세스	중복제거	• 원시데이터를 원천데이터로 정제하는 과정 작성
	비식별화	
데이터 정제 도구	중복제거	• 원시데이터를 원천데이터로 정제하기 위한 도구 및 정제 방안 작성
	비식별화	
데이터 저장 방안	• 정제된 원천데이터를 저장하는 방법 작성	
데이터 관리 방안	• 정제된 원천데이터를 대상으로 품질 관리하는 방안 작성 • 원천데이터 중 기준에 만족하지 못하는 경우 삭제하거나 재 정제하는 방안 작성	

〈 작성요령 〉

- 본 사업에서 정의하는 원천데이터의 의미를 파악하고, RFP의 기준을 만족하도록 '정제', '저장', '관리'하는 방안을 작성한다.
- 라벨링 단계에 들어가기 전에 학습용 데이터로 적합한 데이터를 선별하고, 처리하는 정제 프로세스를 정의한다.
- 데이터 정제는 도구(소프트웨어)를 활용하여 정해진 규칙에 따라 제외 또는 변환하는 방법, 작업자가 직접 눈으로 확인하는 검사하는 방법 등을 적용할 수 있다.
- 작업자에게 제공할 '정제 가이드라인'에 대한 작성 방안을 기재한다.
- [참고] 『제1권. 품질관리 안내서』 내, 'Ⅲ-제2장'

2.4 라벨링 유형

2.4.1 라벨링데이터 정의

- 작성 양식 (예시)

〈표 Ⅳ-8〉 라벨링데이터 정의

구분	내용
가공 담당 기업	• 가공 담당 기업명 작성
라벨링데이터 포맷	• 원천데이터 파일 형식 및 라벨링데이터 파일 형식 작성
라벨링데이터 규모	• 원천데이터 포맷 별 라벨링데이터 구축 규모 작성
라벨링 타입	• 원천데이터를 대상으로 라벨링하고자 하는 어노테이션 방식 작성 • 예시) 이미지 데이터의 경우 'Bounding Box', 'Polygon' 등이 있음

〈 작성요령 〉

- 본 사업에서 정의하는 라벨링데이터의 의미를 파악하고, RFP의 기준을 만족하도록 목표를 정의하여 작성한다.
- 여러 포맷으로 원천데이터가 확보되는 경우, 각각의 포맷 별 라벨링데이터 규모를 상세히 작성한다. (n개의 원천데이터 : n개의 라벨링데이터)
- 하나의 원천데이터를 대상으로, 다양한 라벨링데이터가 구축되는 경우 각각의 라벨링데이터 규모를 상세히 작성한다. (1개의 원천데이터 : n개의 라벨링데이터)
- [참고] 『제1권. 품질관리 안내서』 내, 'Ⅲ-제2장', 'Ⅴ-제2장-2절'

2.4.2 가공방안

● 작성 양식 (예시)

〈표 Ⅳ-9〉 데이터 가공방안

구분	내용
가공 프로세스	• 원천데이터를 대상으로 가공하는 과정 작성
데이터 가공 도구	• 원천데이터를 가공하기 위한 도구 및 가공 방안 작성 • 만약 어노테이션 방식이 다양한 경우, 각각의 가공 도구 및 방안 별도로 작성
데이터 저장 방안	• 가공된 라벨링데이터 및 원천데이터를 저장하는 방법 작성
데이터 관리 방안	• 라벨링데이터를 대상으로 품질 관리하는 방안 작성 • 라벨링데이터 중 기준에 만족하지 못하는 경우 삭제하거나 재 가공하는 방안 작성

〈 작성요령 〉

- 본 사업에서 정의하는 라벨링데이터의 의미를 파악하고, RFP의 기준을 만족하도록 '가공', '저장', '관리'하는 방안을 작성한다.
- 어노테이션 포맷 및 저장 형식, 저장 구조에 대한 내용을 구체적으로 제시한다.
- 데이터 구축 목적 달성을 위해 원천 데이터 형태, 구축 목적에 부합하는 라벨링 도구를 선정하고 세부적인 내용을 제시한다.
- 작업자에게 제공할 '가공 가이드라인'에 대한 작성 방안을 기재한다.
- [참고] 『제1권. 품질관리 안내서』 내, 'Ⅲ-제2장'

2.5 모델 적용 방안

2.5.1 학습모델 임무(TASK) 정의

〈표 Ⅳ-10〉 학습모델 TASK 정의

<table>
<tr><th>구분</th><th colspan="6">내용</th></tr>
<tr><td>모델링 담당 기업</td><td colspan="6">• 학습모델 담당 기업명 작성</td></tr>
<tr><td>TASK 정의</td><td colspan="6">• 선정한 TASK에 대한 정의</td></tr>
<tr><td rowspan="3">TASK 유형 선정</td><td colspan="6">• 아래 5개의 TASK 유형 중 해당되는 유형을 선택하고, 예시에 작성되지 않은 학습모델 TASK 유형의 경우 추가 기재</td></tr>
<tr><td>① 분류</td><td>② 탐지</td><td>③ 추정</td><td>④ 이해</td><td>⑤ 합성</td><td>⑥ 기타</td></tr>
<tr><td>☐</td><td>☐</td><td>☐</td><td>☐</td><td>☐</td><td>☐</td></tr>
<tr><td>TASK 선정 사유</td><td colspan="6">• 선정한 TASK에 대한 선정 사유 작성</td></tr>
</table>

〈 작성요령 〉

- 본 사업에서 정의하는 학습모델 TASK의 의미를 파악하고, RFP의 기준을 만족하도록 목표를 정의하여 작성한다.
- [참고] 『제1권. 품질관리 안내서』 내, 'Ⅳ-제4장', 'Ⅴ-제2장-2절'
- [참고] Paperswithcode.com 내, 'State-of-the-Art', 'Datasets', 'Methods'

2.5.2 학습모델 후보군 선정

〈표 Ⅳ-11〉 학습모델 후보군 정의

<table>
<tr><th>구분</th><th colspan="2">내용</th></tr>
<tr><td>선행 연구</td><td colspan="2">• 선정한 TASK 관련 선행 연구(최신 학습모델 및 알고리즘) 분석내용 기재</td></tr>
<tr><td rowspan="5">학습모델 후보군
(최소2개,
최대5개)</td><td>후보1</td><td>• 학습모델 및 해당 알고리즘 작성
• 예시) '탐지' Task의 경우 'EfficientDet', 'YOLO' 등이 있음</td></tr>
<tr><td>후보2</td><td>• 학습모델 및 해당 알고리즘 작성</td></tr>
<tr><td>후보3</td><td>• 학습모델 및 해당 알고리즘 작성</td></tr>
<tr><td>후보4</td><td>• 학습모델 및 해당 알고리즘 작성</td></tr>
<tr><td>후보5</td><td>• 학습모델 및 해당 알고리즘 작성</td></tr>
</table>

〈 작성요령 〉

- 본 사업에서 정의하는 학습모델 TASK의 의미를 파악하고, RFP의 기준을 만족하도록 목표를 정의하여 작성한다.
- [참고] 『제1권. 품질관리 안내서』 내, 'Ⅳ-제4장', 'Ⅴ-제2장-2절'
- [참고] Paperswithcode.com 내, 'State-of-the-Art', 'Datasets', 'Methods'

2.5.3 학습모델 품질지표 선정

〈표 Ⅳ-12〉 학습모델 품질지표 선정

구분	내용	
학습모델 후보1 유효성 품질관리	품질지표	• 학습모델 후보1에 대한 품질지표 작성 • 예시) '탐지' Task의 경우 'mAP', 'F1-Score' 등이 있음
	선행연구	• 작성한 품질지표에 따른 최신 연구 분석내용 기재 • 예시) 'COCO minival Benchmark'에서 2021년 'Florence-CoSwin-H' 모델이 mAP 62점을 기록
	지표기준	• 작성한 품질지표에 따른 정량적 목표 작성 • 예시) 'mAP' 60이상
학습모델 후보2 유효성 품질관리	품질지표	
	선행연구	
	지표기준	
학습모델 후보3 유효성 품질관리	품질지표	
	선행연구	
	지표기준	
학습모델 후보4 유효성 품질관리	품질지표	
	선행연구	
	지표기준	
학습모델 후보5 유효성 품질관리	품질지표	
	선행연구	
	지표기준	

〈 작성요령 〉

- 본 사업에서 정의하는 학습모델 유효성 평가의 의미를 파악하고, RFP의 기준을 만족하도록 목표를 정의하여 작성한다.
- [참고] 『제1권. 품질관리 안내서』 내, 'Ⅳ-제4장', 'Ⅴ-제2장-3절'
- [참고] Paperswithcode.com 내, 'State-of-the-Art', 'Datasets', 'Methods'

3 1-Cycle 진행방안

3.1 1-Cycle 수행 개요

● 작성 양식 (예시)

〈표 Ⅳ-13〉 1-Cycle 수행 개요

구분	내용
수행 개요	• 1-Cycle 수행에 대한 개념 작성
수행 목적	• 1-Cycle 수행에 대한 목적 작성

〈 작성요령 〉

• 인공지능 학습용 데이터 구축 시 1-Cycle 수행이 가지는 개념 및 목적을 작성한다.
• [참고] 『제1권. 품질관리 안내서』 내, 'Ⅰ-제3장', 'Ⅱ-제2장', 'Ⅴ-제2장'

3.2 계획단계 1-Cycle

3.2.1 계획단계 1-Cycle 수행용 샘플링 데이터 정의

● 작성 양식 (예시)

〈표 Ⅳ-14〉 계획단계 1-Cycle 수행용 샘플링 데이터 정의

구분	내용
목표 수량	• 계획단계 1-Cycle을 수행하기 위한 최소 수량을 기재
샘플링데이터 수집방안	• 과제를 계획하는 단계에서, 빠르게 샘플링데이터를 수집할 수 있는 방안 작성
샘플링데이터 정제방안	• 과제를 계획하는 단계에서, 빠르게 정제할 수 있는 방안 작성
샘플링데이터 가공방안	• 과제를 계획하는 단계에서, 빠르게 가공할 수 있는 방안 작성
샘플링데이터 모델 적용 방안	• 과제를 계획하는 단계에서, 빠르게 학습모델에 적용하여 품질지표 결과를 확인할 수 있는 방안 작성

〈 작성요령 〉

• 계획단계에 적용할 1-Cycle 수행용 샘플링 데이터를 정의한다.
• [참고] 『제1권. 품질관리 안내서』 내, 'Ⅰ-제3장', 'Ⅱ-제2장', 'Ⅴ-제2장'

3.2.2 계획단계 1-Cycle 수행 시기 및 산출물

● 작성 양식 (예시)

〈표 Ⅳ-15〉 계획단계 1-Cycle 수행 시기 및 산출물

시기	수행내용	산출물
W+1	• 샘플링데이터를 수집	• 원시데이터(샘플) • 수집 주의사항 정리표
W+2	• 수집된 샘플링데이터 정제	• 원천데이터(샘플) • 정제 주의사항 정리표
W+3	• 정제된 데이터 가공	• 라벨링데이터(샘플) • 가공 주의사항 정리표
W+4	• 학습모델 적용 및 결과 도출	• 학습모델 유효성 테스트 결과서 • 피드백 점검표

〈 작성요령 〉

• 계획단계에 적용할 1-Cycle 수행에 대한 상세 내용 및 산출물을 시기별로 작성한다.
• 실행단계에서 발생할 수 있는 품질관리 이슈를 최소화하기 위한 방향으로, 각각의 산출물을 정의한다.
• 시기를 작성할 때는 사업자 선정 또는 우선협상대상자 선정을 기점으로 W(Week, 주별) 단위로 계산한다.
• [참고] 『제1권. 품질관리 안내서』 내, 'Ⅰ-제3장', 'Ⅱ-제2장', 'Ⅴ-제2장'

3.3 실행단계 1-Cycle

3.3.1 초기데이터 1-Cycle 수행용 데이터 정의

● 작성 양식 (예시)

〈표 Ⅳ-16〉 실행단계 1-Cycle 수행용 초기데이터 정의

구분	내용
목표 수량	• 초기데이터 1-Cycle을 수행하기 위한 최소 수량을 기재
목표 일정	• 초기데이터 1-Cycle을 최종 완료하는데 필요한 시간을 산정하여 상세히 작성
피드백 방안 (계획 → 초기)	• 계획단계 1-Cycle 수행 결과를 기반으로, 실행단계에 적용하기 위한 방안 작성
피드백 방안 (초기 → 실행)	• 초기데이터 1-Cycle 수행 결과를 기반으로, 향후 실행단계에 적용하기 위한 방안 작성

〈 작성요령 〉

• 실행단계에 적용할 1-Cycle 수행용 초기데이터를 정의한다. (일반적으로 30% 이내)
• 1-Cycle 수행 결과를 피드백 하는 방안에 대해 기재한다.
• [참고] 『제1권. 품질관리 안내서』 내, 'Ⅰ-제3장', 'Ⅱ-제2장', 'Ⅴ-제2장'

3.3.2 초기데이터 1-Cycle 수행 시기 및 산출물

- 작성 양식 (예시)

〈표 Ⅳ-17〉 초기데이터 1-Cycle 수행 시기 및 산출물

시기	수행내용	산출물
M+0.5	• 원시데이터를 수집	• 원시데이터 • 수집 가이드라인
M+1.0	• 수집된 원시데이터 정제	• 원천데이터 • 정제 가이드라인
M+1.5	• 정제된 원천데이터 가공	• 라벨링데이터 • 가공 가이드라인
M+2.0	• 가공된 라벨링데이터 학습모델 적용 및 결과 도출	• 학습모델 유효성 테스트 결과서 • 피드백 점검표

〈 작성요령 〉

- 초기데이터 1-Cycle 수행에 대한 상세 내용 및 산출물을 시기별로 작성한다.
- 향후 실행단계에서 발생할 수 있는 품질관리 이슈를 최소화하기 위한 방향으로, 각각의 산출물을 정의한다.
- 시기를 작성할 때는 사업자 선정 또는 우선협상대상자 선정을 기점으로 M(Month, 월별) 단위로 계산한다.
- [참고] 『제1권. 품질관리 안내서』 내, 'Ⅰ-제3장', 'Ⅱ-제2장', 'Ⅴ-제2장'

4 작업자 운영방안

4.1 작업자 가이드 작성 방안

● 작성 양식 (예시)

〈표 Ⅳ-18〉 작업 가이드라인(안)

구분	내용
작업 조건	• 작업자들이 오해하지 않도록 작업 기준에 따라 작업 조건을 작성
작업 수량	• 작업 수량의 경우 정확한 단위에 따른 기준을 명시해야 함
작업 예시 설명	• 반드시 관련 이미지와 함께 제시되어 작업자과 이해할 수 있도록 해야 함
도구 조작 방식	• 설치가 필요한 경우 설치 파일을 다운로드 받는 것부터 시작해서 설치하고 환경을 설정하고 사용하는 화면 순으로 실제 작업자의 작업 순서에 따라 설명
반려 조건	• 올바른 작업 기준을 준수하기 위해 작업할 수 있도록 작업 결과물에 대한 검사 결과 반려 기준을 반려 작업에 대한 설명과 이미지를 통해서 명확하게 제시
작업 불가 조건	• 작업 대상물에 작업 기준에 해당하는 객체가 존재하지 않거나 작업 방법으로 작업이 불가능한 경우를 명확하게 제시
기타 주의사항	• 작업자들이 작업할 때 작업 환경이나 도구 등에서 주의가 필요한 경우 반드시 작업가이드 해당 내용을 명시하고 가능하면 사전 교육을 통하여 주의사항을 명확하게 전달해야 함

〈 작성요령 〉

• 작업자(크라우드워커)를 대상으로 제공하는 작업 가이드라인 초안을 작성한다.
• [참고] 『제2권. 데이터구축 안내서』 내, 'Ⅰ-제2장'

4.2 작업 할당 방안

- 작성 양식 (예시)

〈표 Ⅳ-19〉 작업 할당 방안

구분	내용
1인당 작업 건수	• 작업 설계 후 직접 수행해 본 결과를 바탕으로 산출하되, 작업 최소 단위로 구분해야 하며, 1시간 기준 작업 및 검수 가능 건수가 최저 시급보다 높도록 설계
1인당 최대 건수	• 작업 / 검수 가능한 최대 수치를 확인하여 인력별 숙련도에 소요되는 시간을 확인하고, 1인당 최대 작업 및 검수 가능 평균 건수를 산출
추가인력 투입방안	• 전체 작업 및 검수 건수를 관리가 가능한 수준에서 확대해 가면서 진행하되 부정 작업자나 불량 작업자에 대한 필터링을 함께 하면서 추가 투입된 작업자 및 검수자에 대한 진척률 및 불량률을 모니터링 함
프로젝트 운영방안	• 여러 가지 사유로 인한 크라우드워커의 이탈을 고려한 프로젝트 운영방안 제시

〈 작성요령 〉

• 작업자(크라우드워커)의 작업 할당 기준을 작성하고, 프로젝트 운영 방안을 제시한다.
• [참고] 『제2권. 데이터구축 안내서』 내, ‘Ⅰ-제2장’, ‘Ⅰ-제3장’

4.3 작업 모니터링 방안

● 작성 양식 (예시)

〈표 Ⅳ-20〉 작업 모니터링 방안

구분	지표	내용
작업량 모니터링	작업 건수	• 해당 모니터링 방안 기재
	작업 소요 시간	• 해당 모니터링 방안 기재
	작업자수	• 해당 모니터링 방안 기재
	검수 대기 건수	• 해당 모니터링 방안 기재
	검수 완료 건수	• 해당 모니터링 방안 기재
	검수 소요 시간	• 해당 모니터링 방안 기재
	검수자수	• 해당 모니터링 방안 기재
진척률 모니터링	1일 평균 작업/검수 건수	• 해당 모니터링 방안 기재
	작업/검수 소요시간	• 해당 모니터링 방안 기재
	1인 작업/검수 건수	• 해당 모니터링 방안 기재
	예상 완료율	• 해당 모니터링 방안 기재
불량률 모니터링	전체 반려 건수/반려율	• 해당 모니터링 방안 기재
	반려가 많은 작업자	• 해당 모니터링 방안 기재
	반려가 적은 검수자	• 해당 모니터링 방안 기재
기타 사항	• 기타 모니터링 방안 기재	

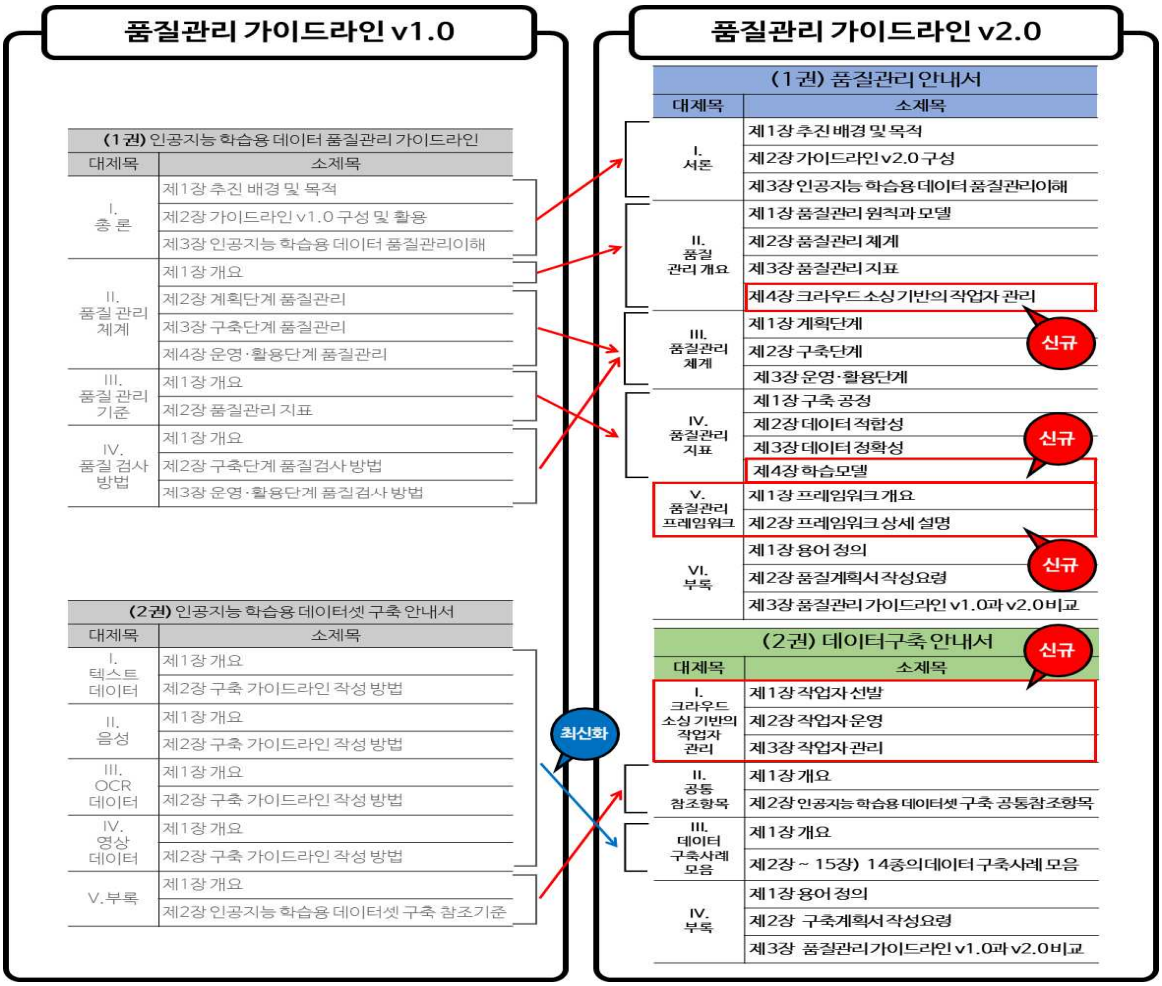
〈 작성요령 〉

• 작업자(크라우드워커)의 작업내용 모니터링 방안을 작성한다.
• [참고] 『제2권. 데이터구축 안내서』 내, ‘Ⅰ-제2장’, ‘Ⅰ-제3장’

제3장 품질관리 가이드라인 v1.0과 v2.0 비교

- ‘품질관리 가이드라인 v2.0’은 2020년도 구축 데이터 170종과 2021년도 구축 데이터 190종의 유형과 현황을 분석한 자료를 대상으로 진행한 품질관리 관련 각종 고도화 작업의 결과(품질관리 프레임워크 기반 마련 등)를 반영하였다. 아래 그림에서는 2021년 2월에 발간된 ‘품질관리 가이드라인 v1.0’과의 내용 구성 부분의 차이에 대한 이해를 돕기 위해 비교하여 제시하였다.

※ 품질관리 가이드라인 v1.0과 v2.0은 모두 ‘AI 허브’에 게시되어 있으므로 다운로드 받아 활용 할 수 있다.



[그림 Ⅳ-1] 품질관리 가이드라인 v1.0과 v2.0 비교

제4장 참고자료

1. 한국정보통신기술협회(TTA), AI 학습용 데이터 구축사업 공통기준
2. 한국정보통신기술협회(TTA), 2020년 인공지능 학습용 데이터 구축사업 최종산출물 검토
3. 한국지능정보사회진흥원(NIA), 2021년 인공지능 학습용 데이터 구축사업 사업수행계획서 검토
4. 한국정보통신기술협회(TTA), 2021년 인공지능 학습용 데이터 구축사업 품질검증합의서 검토
5. AI Hub (https://aihub.or.kr/) 각 구축사업 별 산출물 검토

품질관리 가이드라인 v2.0

2022년 2월 발행

발행처: 한국지능정보사회진흥원 (NIA)

〈 『가이드라인』 작성 참여진 〉

한국지능정보사회진흥원 고윤석 본부장

한국지능정보사회진흥원 오현목 팀장

한국지능정보사회진흥원 유호진 팀장

한국지능정보사회진흥원 박수인 수석

한국지능정보사회진흥원 지능데이터본부 품질유닛
(김진호 수석, 최인언 선임, 홍현우 주임, 윤주미 연구원, 김지수 연구원)

크라우드웍스 박영진 본부장

에스에스엘 이태석 팀장

아이피투비 주희엽 대표

〈 자문 위원 〉

과학기술연합대학원대학교 성원경 교수

광주과학기술원 주일택 교수

국립암센터 김대홍 박사

서울시립대학교 이재호 교수

세종대학교 구영현 교수

한국광기술원 정효영 박사

비오피 조용현 부대표

써로마인드 장하영 대표

자이플래닛 유석 부사장

지농 박흔동 대표

• 본『데이터구축 안내서』는 지능정보산업 인프라 조성을 위한 인공지능 학습용 데이터 구축사업 중 '인공지능 학습용 데이터 품질관리체계 고도화 및 전문기술지원 컨설팅' 용역 사업의 결과 산출물입니다.
• 한국지능정보사회진흥원 지능데이터본부 품질유닛 adp_qa@nia.or.kr

제2권

데이터구축 안내서

과학기술정보통신부

NIA 한국지능정보사회진흥원