

인공지능 데이터 구축 가이드라인

- 113. 태풍 및 홍수로 인한 피해 및 위험 데이터 -

담당 역할	기관명
데이터 구축 총괄	스마트쿵㈜
데이터 설계	스마트쿵㈜
데이터수집 및 정제	(주)엘씨씨코리아
가공(라벨링, 어노테이션)	(주)인더스웰, 스마트쿵㈜
데이터 검수(자체 검수)	디노플러스㈜

구축 가이드라인 작성	스마트쿵㈜	최진욱
	(주)인더스웰	이상기
가이드라인 버전 (제작일자)	ver 1.4 ('22.02.28)	

목 차

1. 데이터 구축 개요	1
2. 문제 정의	3
2.1 임무 정의	3
2.2 데이터 구축 유의사항	5
3. 데이터 수집·정제	6
3.1 원시데이터 선정	6
3.2 수집·정제 절차 및 방법	7
3.3 수집·정제 기준	10
3.4 수집·정제 조직	17
3.5 수집·정제 도구	19
4. 데이터 가공	21
4.1 가공 절차 및 방법	21
4.2 가공 기준	23
4.3 가공 규격	28
4.4 가공 조직	32
4.5 가공 도구	33
5. 검수	34
5.1 검수 절차 및 방법	34
5.2 검수 기준	35
5.3 검수 조직	38
5.4 검수 도구	41
5.5 기타 품질관리 활동	42

1. 데이터 구축 개요

□ 구축배경 및 필요성

- 홍수 및 태풍에 의한 홍수는 이상기온으로 매년 증가하는 추세에 있으며, 강우량도 지금까지 경험해보지 못한 규모로 증가추세에 있음
- 태풍 및 홍수로 인한 피해 및 위험데이터를 수집하고 구축하여 피해 규모에 따른 복구를 지원해야 될 필요성이 있음
- 태풍에 의한 국지성 호우, 홍수 등으로 인한 폐기물 부유, 침수, 오염 등 피해와 위험 요소(침수, 산사태 등의 위험)에 대해 신속한 복구를 위한 데이터 구축이 필요함

□ 구축방향 및 목표

- 태풍 및 홍수로 인한 폐기물 부유, 침수, 오염 등 피해 및 위험 데이터를 구축하여 향후 피해 규모를 줄이기 위한 선제적 제거와 피해 규모에 따른 신속한 복구를 지원하기 위한 인공지능 학습용 데이터셋을 구축함
- 이를 위하여 태풍 및 홍수로 인한 피해 상황에 대하여 하천, 도시 등의 지역에서 교량 및 교각, 교통신호등, 전신주, 조경수, 도로 등의 침수 상황 및 부유물·잔존물 카테고리를 포함한 16종의 데이터를 구축하는 것을 목표로 하며, 정상상태 18만장, 재해상태 12만장 이상의 데이터를 구축함

□ 데이터 구축 프로세스

- 데이터 수집 전제조건



- 이미지 수집 및 추출 가능 장치를 통한 데이터 수집

- 디지털 카메라(스마트 폰) : 댐하류의 정상상태/홍수상태 데이터 및 댐 상류의 정상상태 데이터의 획득에 사용
- 드론 : 홍수상태에서 접근이 어려운 댐 상류 지역이나 국지성 호우로 접근이 불가능한 지역의 데이터 획득에 사용

- 데이터 취득 장소 : 하천 (전국 11개 수계에 27개의 댐의 상류와 하류), 도시 (호수 및 태풍 발생 도시지역)
- 장소별(세부 장소구분), 시간별, 날씨별 데이터 수집
- 홍수 및 태풍의 시점이 정확하지 않기 때문에 이에 대한 예측 시스템(국내·외 기상청 정보 등)을 활용해 데이터 수집 대비태세를 수립

2. 문제 정의

2.1. 임무 정의

□ 데이터 분석 정의

- 구축 데이터 분석 방법은 시각화, 공간분석, 자료 분석, 통계분석, 데이터마이닝 등 다양한 분석 방법을 통해 분석 가능
- 본 데이터는 태풍 및 홍수로 인한 피해 및 위험 데이터의 수집이 필요 이에 따라서 홍수 및 태풍에 취약한 하천 과 도시의 두 대분류를 통해 구분해 수집
- (하천분류) : 하천홍수의 경우, 댐의 상류지역과 하류지역을 구분하여 데이터를 획득할 예정이며 전국 11개의 수계에 설치되어 있는 27개의 댐의 상류와 하류를 구분하여 데이터를 획득하는데, 이는 댐 상류지역과 하류지역이 관리기관이 다르기 때문
- (도시 분류) 도시홍수는 전국 도시별로 태풍, 집중호우 등으로 예측불가하기 때문에 발생시점에 집중적으로 수집하는 것으로 하나, 국가의 재해발생 예상수치를 참고로 하여 재해발생빈도가 가장 높은 지역을 중심으로 비상대비체제를 준비

□ 분석 방법

- 기 발생된 홍수 데이터의 수집
 - 기 발생된 자연재해 데이터는 2017년도부터 2020년도의 데이터는 환경청에서 조사한 홍수피해가 심한 지역(홍수 흔적 자료 조사 지역)을 대상으로 선정
 - 2016년 이전에는 정부 보고서가 없는 관계로 정부에서 발생한 각 년도별 재해 연보(행안부)를 참고

구분	자료명/자료 종류	발행처/획득 장소
홍수 영상 데이터	홍수 피해 발간 조사 보고서(2017 ~ 2020)	환경부
	홍수 피해 연보(2016년 이전)	행안부
수문 데이터	홍수위 관련 데이터 수문개방/방류량 데이터	홍수 통제소
	강우량 관련 데이터	기상청
기타	WAMIS(국가수자원관리종합정보시스템) 강수량 자료 수위자료 기상자료 댐수문자료 홍수량 자료 홍수위험지도정보 기후변화 국가 재난 관리 시스템	환경부

○ 2020년도 홍수 데이터의 수집 대상

- 2020년의 주요 대상 범위는 2020년도 홍수 피해 대상 271개의 하천 홍수 중 가장 피해가 심한

34개 지역을 중심으로 조사함

- 2017~2020년까지 최근 4년간 홍수 피해가 가장 심한 강원, 경기, 충북 등 지역을 위주로 데이터 수집을 추진

□ 구축계획

- 데이터는 총 300,000장을 구축하며, 정상데이터와 피해데이터의 비율은 6:4로, 각 180,000장, 120,000장을 수집함
- 원칙적으로 재해상태 데이터에 대조되는 정상상태 데이터는 해당 장소로 가서 재수집하는 것을 우선으로 하되, 인공지능 학습용 데이터로 활용되는 범위에서 유사 데이터로 대체함

(단위 : 건)

구분		구축계획		
카테고리	객체	합계	정상	피해
하천	댐	5,000	3,000	2,000
	제방(운동기구/놀이기구/기타)	55,000	33,000	22,000
	교량 및 교각	15,000	9,000	6,000
	소계	75,000	45,000	30,000
도시	교통신호등	15,000	9,000	6,000
	보행신호등	20,000	12,000	8,000
	교통안전표지판(소형)	15,000	9,000	6,000
	이정표표지판(대형)	15,000	9,000	6,000
	사람	15,000	9,000	6,000
	자동차	20,000	12,000	8,000
	산사태	5,000	3,000	2,000
	도로	15,000	9,000	6,000
	건물	20,000	12,000	8,000
	소계	140,000	84,000	56,000
공통·기타	가로등	25,000	15,000	10,000
	조경수	30,000	18,000	12,000
	전신주	25,000	15,000	10,000
	지하차도	5,000	3,000	2,000
	소계	85,000	51,000	34,000
구축합계		300,000	180,000	120,000

2.2. 데이터 구축 유의사항

☐ 개인정보

- 원칙적으로 사람이 있는 데이터는 수집하지 않음
- 사람이 객체인 경우 비식별화 처리를 통해 개인식별이 불가능하게 함.

☐ 민감정보

- 전화번호 및 자동차번호판 등도 개인정보에 준하는 비식별화 처리를 통해 정보식별이 불가능하게 함.
- 번호판의 일부 노출의 경우 전체 차량번호를 추정할 수 없는 수준의 경우에는 일부 허용함

☐ 저작권정보

- 수집자 및 수집기관의 직접촬영을 통해 데이터를 수집하는 것을 원칙으로 함
- 구매 데이터의 경우 저작권 협의를 통해 재가공된 데이터의 경우 공개 가능함을 확인함

☐ 공개원칙

- CCL 표기법으로 BY-NC-SA(저작물표시-비영리-동일조건변경허락) 형태로 공개함

☐ 촬영사항

- 자격 : 드론을 영리목적으로 운영하지 않았으므로, 본 사업에서는 특정한 자격을 요하지 않음
- 사고 및 책임보험 : 클라우드 소싱으로 진행하여 사업 측면에서 보험 등을 고려하지 않음
- 주요 인력 : 클라우드 워커 및 재해 발생 지역 인근 주민 등 일반인, 과제참여 인력 등
- 안전검점 : 수집기관별 자체 안전 점검을 통해 촬영함
- 보안지역 : 보안지역 촬영은 원칙적으로 진행하지 않음

3. 수집·정제

3.1. 원시데이터 선정

☐ 수집

○ 이미지 데이터

- 클라우드 소싱 수집 : 180,000장
- 구매 데이터 : 언론사 데이터 구매 120,000장

○ 동영상 데이터

- 드론촬영 동영상 데이터를 활용 초당 30프레임 기준 1장을 추출하여 사용함

○ 정보처리

☐ AI 학습 정보

○ 데이터 포맷 : jpg 및 json

○ 재해 상황에 대한 라벨링은 Bounding Box 및 Polygon 정보를 포함함

☐ AI 학습에 충분한 종수, 사례 및 대표성

○ 재해 유형별로 최소 1,000장 이상으로 종수별 데이터 편중을 방지하여 수집함

○ 태풍 및 호우로 인한 자연재해에 한정되어 피해지역을 기준으로 하천, 도시, 공통 등 3개 지역을 선정함

○ 개인정보보호와 관련한 비식별화 처리를 진행하여 가공함

3.2. 수집·정제 절차 및 방법

□ 수집·정제 절차

○ 수집 절차

- (선정) 피해발생 예상지역에 대한 건축구조전문가의 사전점검 및 예상 피해 시설확인
- (점검) 데이터 수집의 구체적인 방법에 대한 점검 (촬영각도, 거리 등 필요사항 확인)
- (교육) 클라우드 소싱 인력에 대한 데이터 수집 가이드 교육 및 실습
 - 온·오프라인 교육 병행 운영
- (수집) 가이드라인에 의한 데이터 수집
 - 사전에 협의된 카테고리 폴더별로 클라우드 또는 NAS에 업로드
- (검수) 수집 가이드라인 적합 여부 검수
 - 적합 : 정제(비식별화 등) 단계로 이관
 - 부적합 : 부적합 사유를 기재하여 수집기관과 공유 및 재촬영 요청

○ 정제 절차

- (비식별화) 사람 얼굴, 전화번호, 자동차번호판, 주소, 이름, GPS 정보 등 개인 식별, 지역 특성이 가능한 정보에 대한 비식별화 진행
- (검수) 비식별화에 대한 전량 전수 검사 후 가공 단계로 이관

□ 데이터 획득 절차

- 집중적인 호우 피해가 예상되는 지역을 선정
- 과제 수행자(건축구조기술사)가 그 지역을 방문하여, 피해가 예상되는 시설물들을 추출하고, 경험적이고 계산적인 측면에서 그 이유를 개략적으로 확인
- 과제 수행자에 의해 기본적인 사전조사가 끝나면 클라우드워커 인원을 모집
- 사진촬영 방법 및 사전조사 대상에 대하여 숙지하고 매뉴얼대로 정지 이미지 데이터를 수집하고, 위치도를 작성(정지이미지 데이터는 각 피사체 별로 사면전경 사진 4매, 접지, 접합부 사진 각 1매)
- 태풍 및 호우 등 자연재해가 발생한 후 동일한 지역을 다시 방문하여 피해상황을 동일한 조건에서 촬영
- 자연재해 데이터는 수집기관에서 기 보유한 피해 데이터를 우선 검수하고 부족 시 재해 데이터 구매를 추진함
- 데이터 추가 확보 필요 시 언론사, 방송국 등으로 부터 데이터 구입
- 자연재해 피해 전후 상황을 비교하고 어노테이션 하여 모델링
- 시설물 피해발생 원인분석 전문가인 과제수행자(건축구조기술사)가 함께하여 정확한 피해 원인을 어노테이션 하고, 이를 인공지능 학습에 반영

[데이터 획득·정제 방안]

	데이터 획득 형태	수집장비	데이터 형식	데이터 처리	담당 인원
1	야외현장 촬영	디지털카메라 (모델명 : 000)	RAW → PNG → JPG	데이터 수집 및 라벨링 툴	과제 인력 및 크라우드소싱 인력
2	웹 크롤링	크롤링 서버	여러 이미지 포맷 → JPG	데이터 수집 및 라벨링 툴	외주 크롤링 담당자
3	기존자료	디지털 카메라	여러 이미지 포맷 → JPG	데이터 수집 및 라벨링 툴	과제 인력

□ 수집·정제 방법

○ 수집 방법

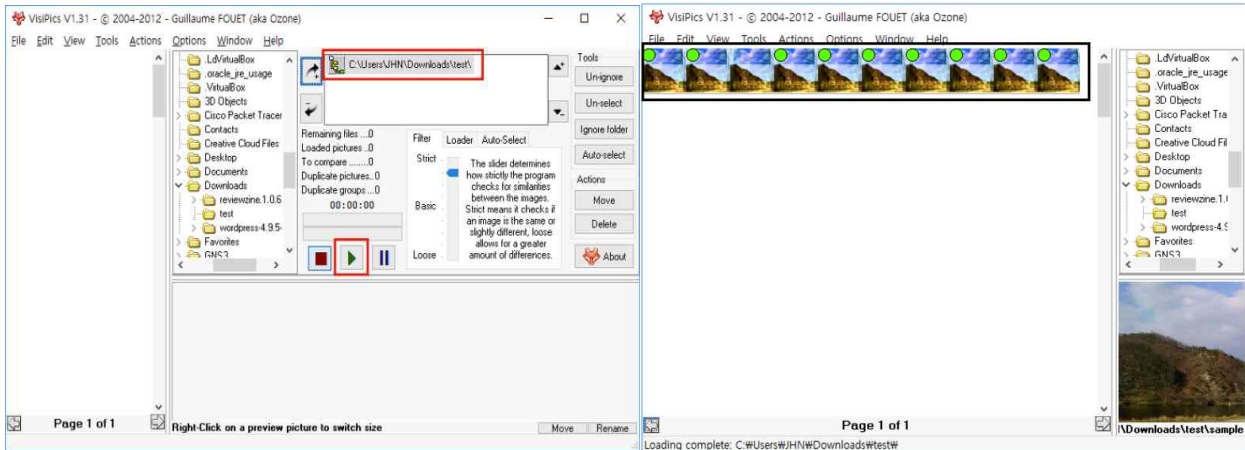
- 직접 촬영 또는 구매
- 데이터 수집 장비
 - 스마트폰(디지털 카메라) : 클라우드 소싱 인력들이 사용하는 장비(촬영날짜와 위치추정 가능)
 - 드론 : 높은 곳 등 사람의 접근이 불가능한 지역의 데이터 획득을 위한 장비
- 데이터 획득 장소
 - 피해 빈발 지역
 - ※ 홍수 위험지도를 기준으로 위험도가 높은 지역을 선정하여 홍수 데이터 취득 계획을 수립함
 - 시설물 밀집도가 높아 피해가 클 것으로 예상되는 서울 등 수도권 지역

○ 정제 방법

- (1차 정제) 자동화툴을 사용한 중복데이터 제거 및 비식별화 등 정제
- (2차 정제) 육안 검수 후 인력 활용 수동 비식별화
- 육안 검수를 통한 전수검사

□ 원천데이터 1차 정제 가이드

- (정의) 1인당 할당될 파일을 구분하고, 원천데이터 품질지표 기반 충족하지 못하는 이미지를 제거하는 작업
- (작업자의 부주의로 인한 데이터 편집) 같은 장소를 반복적으로 촬영한 경우, 흔들린 이미지를 취득한 경우 등
- (방법) 중복검사 프로그램을 활용하여 중복 혹은 유사 이미지 분류 작업 수행
- 육안 작업과 병행하여 실시함



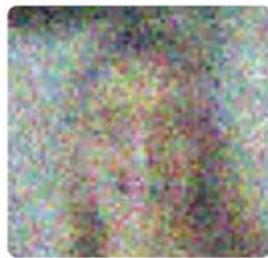
[중복검사 프로그램 사용 예시]

□ 1차 정제 데이터 2차 정제(비식별화) 가이드

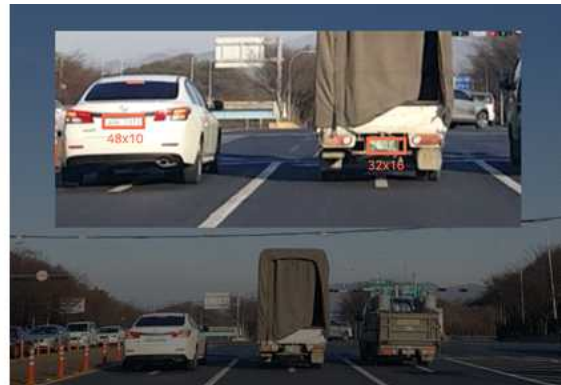
- (정의) 개인정보(얼굴, 자동차 번호판 등) 또는 민감정보(저작권 침해 우려가 있는 자료, 군사시설 등)가 취득된 경우 비식별화 수행
- (비식별화 도구 개요) 본 과제에 사용되는 비식별화 도구를 사용하며, 오토 기능이 탑재되어 있어 사람의 얼굴과 자동차 번호판을 자동으로 비식별화함



원본 데이터



익명화 데이터



[그림. 개인정보 비식별화 대상 및 예시]

- (비식별화 주의사항) 구축된 데이터는 인공지능 학습을 위한 자료로 공개되며, 법적 문제와도 직결되는 사항이기 때문에 꼼꼼한 작업과 철저한 검사가 필요
- 오토 기능을 이용하여 비식별화를 진행하였다 해도 작업자는 이미지를 수동으로 다시 확인하며 미비한 부분을 수작업을 통해 추가 작업을 해야 함

3.3. 수집·정제 기준

□ 수집 기준

○ 다양한 촬영높이와 촬영거리를 기준으로 다량의 이미지를 획득하기 위하여 대상체를 중심으로 촬영 가능한 각도 내에서 이동하며 촬영

- 카메라

- 각 객체의 전면 이미지, 양측면 이미지를 수집(전면 기준으로 15도, 30도)
- 카메라의 GPS 및 촬영 방향값을 관리하여 정상상태 촬영위치와 재해상태 촬영위치의 오차를 최소화함

- 드론

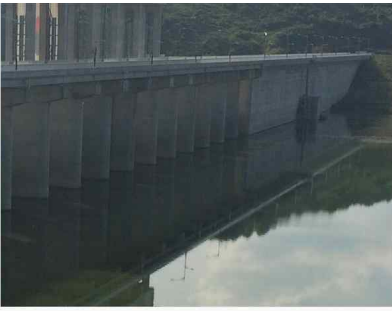
- 카메라로 객체 전체를 인식할 수 없거나 수집이 불가능할 경우 사용
- CCTV가 존재하는 경우 CCTV 이미지를 참고하여 해당 위치에서 드론으로 이미지 수집(사람이 카메라로 찍을 수 없는 위치나 각도의 경우 사용)
- 상공 50m를 기준으로 객체 전체를 인식할 수 있을 만큼 거리를 1m씩 줄이거나 늘려 최대 10m까지 차이나도록 수집(ex: 50m를 기준으로 49, 48, 47... 45m에서 각각 촬영하여 수집)
- 전면뷰(드론의 경우 항공뷰)와 각 측면부(동서남북)의 지면 기준 45도, 60도에서 각 이미지를 수집

분류	항목	필요사항
수집장비	카메라(스마트폰)	Full HD 이상의 화질 촬영가능 카메라(DSLR, 스마트폰 카메라 등)
	드론	Full HD 화질 이상의 카메라 탑재 드론
장소	하천지역	27개의 댐의 상류와 하류
	도시지역	자연재해(홍수 및 태풍) 발생 도시(재해발생빈도 높은 도시 대상 비상대치체제 수립)
날씨	비행날씨	홍수, 태풍, 맑음 또는 재해 미발생 날씨
시간	수집시간	주간, 야간 구분
사람	수질 분석 전문가	홍수 시 하천 혹은 댐의 수질 변화에
	댐 관리 전문가	댐 상류의 촬영을 위한 안내자, 수계 및 댐위치별 지역홍수통제소 담당 직원

○ 클라우드 워커 참고를 위한 객체 파손 레벨에 따른 가이드라인 샘플 예시



댐 피해 Lv1



댐 피해 Lv1



댐 피해 Lv1



다리 피해 Lv1



다리 피해 Lv1



다리 피해 Lv1



건물 피해 Lv1



건물 피해 Lv1



건물 피해 Lv1



산사태 피해 Lv1



산사태 피해 Lv1



산사태 피해 Lv1



자동차 피해 Lv1



자동차 피해 Lv1



자동차 피해 Lv1



사람 Lv1



사람 Lv1



사람 Lv1



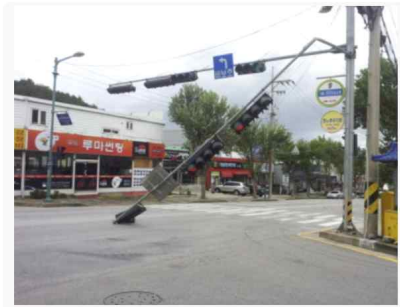
신호등 피해 Lv1



신호등 피해 Lv1



신호등 피해 Lv1



전신주 피해 Lv1



전신주 피해 Lv1



전신주 피해 Lv1





□ 구축 데이터 정의

○ 위험레벨 정의

- 인공지능 학습용 데이터셋 사업의 자연재해 데이터 구축 목적 및 내용에 따라 구축 데이터셋을 "자연재해"분야의 레벨을 정상(Lv.0), 재해(Lv.1)의 2개 레벨로 정의함

구분	레벨	레벨 정의 기준
정상	Lv.0	정상
재해	Lv.1	파손, 손상 및 복구불능

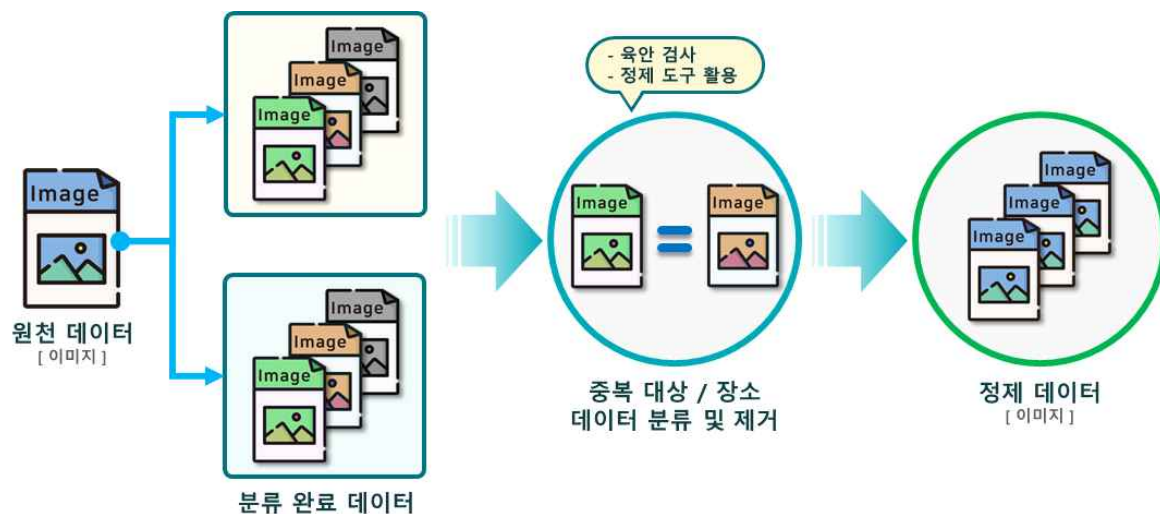
- 구체적인 취약레벨을 정의할 수 없는 객체에 대하여는 취약레벨과 재해레벨을 통합하여 제해레벨로 수집함

재해지역	객체	재해레벨	레벨정의	수량(건)
하천	제방 (운동기구/놀이 기구/기타)	정상 Lv.0	▪ 정상 상태의 제방	55,000
		재해 Lv.1	▪ 하천의 수위상승으로 인한 범람 및 붕괴 ▪ 제방의 하천 수위상승으로 인한 부유물 잔존	
	댐	정상 Lv.0	▪ 정상 상태의 댐	5,000
		재해 Lv.1	▪ 주의, 경계, 심각단계 및 그로 인한 부유 물 생성	
	교량 및 교각	정상 Lv.0	▪ 정상 상태의 교량 및 교각	15,000

재해지역	객체	재해레벨	레벨정의	수량(건)
		재해 Lv.1	<ul style="list-style-type: none"> 하천의 수위상승으로 인한 교각 및 교량의 침수 및 파손, 손상 교각 및 교량에 수위상승으로 인한 부유물 잔존 	
도시	교통신호등	정상 Lv.0	<ul style="list-style-type: none"> 정상 상태의 교통신호등 	15,000
		재해 Lv.1	<ul style="list-style-type: none"> 교통신호등의 침수, 파손, 손상 및 침수로 인한 부유물 잔존 	
	보행신호등	정상 Lv.0	<ul style="list-style-type: none"> 정상 상태의 보행신호등 	20,000
		재해 Lv.1	<ul style="list-style-type: none"> 보행신호등의 침수, 파손, 손상 및 침수로 인한 부유물 잔존 	
	교통안전표지판(소형)	정상 Lv.0	<ul style="list-style-type: none"> 정상 상태의 소형 교통안전표지판 	15,000
		재해 Lv.1	<ul style="list-style-type: none"> 소형 교통안전표지판의 침수, 파손, 손상 및 침수로 인한 부유물 잔존 	
	이정표표지판(대형)	정상 Lv.0	<ul style="list-style-type: none"> 정상 상태의 대형 이정표표지판 	15,000
		재해 Lv.1	<ul style="list-style-type: none"> 대형 이정표표지판의 침수, 파손, 손상 및 침수로 인한 부유물 잔존 	
	사람	정상 Lv.0	<ul style="list-style-type: none"> 정상 상태의 사람 	15,000
		재해 Lv.1	<ul style="list-style-type: none"> 침수로 인해 사람의 발목이상 잠김 	
	자동차	정상 Lv.0	<ul style="list-style-type: none"> 정상 상태의 자동차 	20,000
		재해 Lv.1	<ul style="list-style-type: none"> 자동차 타이어 높이의 1/3이상 침수, 파손, 손상 및 침수로 인한 부유물 잔존 	
	산사태	정상 Lv.0	<ul style="list-style-type: none"> 정상 상태의 산능선, 봉우리 등 	5,000
		재해 Lv.1	<ul style="list-style-type: none"> 산사태 발생 및 산사태로 인한 토사 및 부유물 잔존 	
	도로	정상 Lv.0	<ul style="list-style-type: none"> 정상 상태의 도로(차선 구분 가능) 	15,000
		재해 Lv.1	<ul style="list-style-type: none"> 도로의 침수, 파손, 손상 및 침수로 인한 부유물 잔존 	
	건물	정상 Lv.0	<ul style="list-style-type: none"> 정상 상태의 건물 	20,000
		재해 Lv.1	<ul style="list-style-type: none"> 건물의 침수, 파손, 손상 및 침수로 인한 부유물 잔존 	
공통 및 기타	가로등	정상 Lv.0	<ul style="list-style-type: none"> 정상 상태의 가로등 	25,000
		재해 Lv.1	<ul style="list-style-type: none"> 가로등의 침수, 파손, 손상 및 침수로 인한 부유물 잔존 	
	조경수	정상 Lv.0	<ul style="list-style-type: none"> 정상 상태의 조경수 	30,000
		재해 Lv.1	<ul style="list-style-type: none"> 조경수의 침수, 파손, 손상 및 침수로 인한 부유물 잔존 	
	전신주	정상 Lv.0	<ul style="list-style-type: none"> 정상 상태의 전신주 	25,000

재해지역	객체	재해레벨	레벨정의	수량(건)
		재해 Lv.1	▪ 전신주의 침수, 파손, 손상 및 침수로 인한 부유물 잔존	
	지하차도	정상 Lv.0	▪ 정상 상태의 지하차도	5,000
		재해 Lv.1	▪ 지하차도의 침수, 파손, 손상 및 침수로 인한 부유물 잔존	
총계(건)			300,000	

□ 원시 데이터 정제 프로세스



[그림. 원시 데이터 정제 프로세스]

- 이미지 및 영상수집 장치로부터 획득한 영상데이터, 이미지 데이터는 각 모듈별 메모리, 현장 컴퓨터, 수집서버 3중으로 백업 진행
- 획득된 원시 데이터 관리를 위해 일지와 수집기관별 폴더 형태로 구성하여 원시데이터 저장(네트워크 수집 서버)
- 영상 데이터 정제 프로세스



[그림. 영상 데이터 추출 기준]

- 드론 영상 데이터의 경우 영상 편집 툴을 이용해 단위 시간당 프레임을 추출하여 이미지 데이터로 만든 뒤 정제 작업 실시
- 원천데이터의 정제(비식별화) 절차 : 1차 편집된 원천데이터(이미지)를 비식별화 도구를 통해 자동 비식별화 처리 실시

- 비식별화 완료 후 작업자가 비식별화 처리가 잘되었는지 모든 이미지 검수 및 수정 진행함
- 정면이 아닌 특정 각도에서는 얼굴 인식이 어려울 수 있으며, 문자나 숫자형식의 개인정보는 반드시 작업자가 직접 영역을 지정하여 비식별화 처리

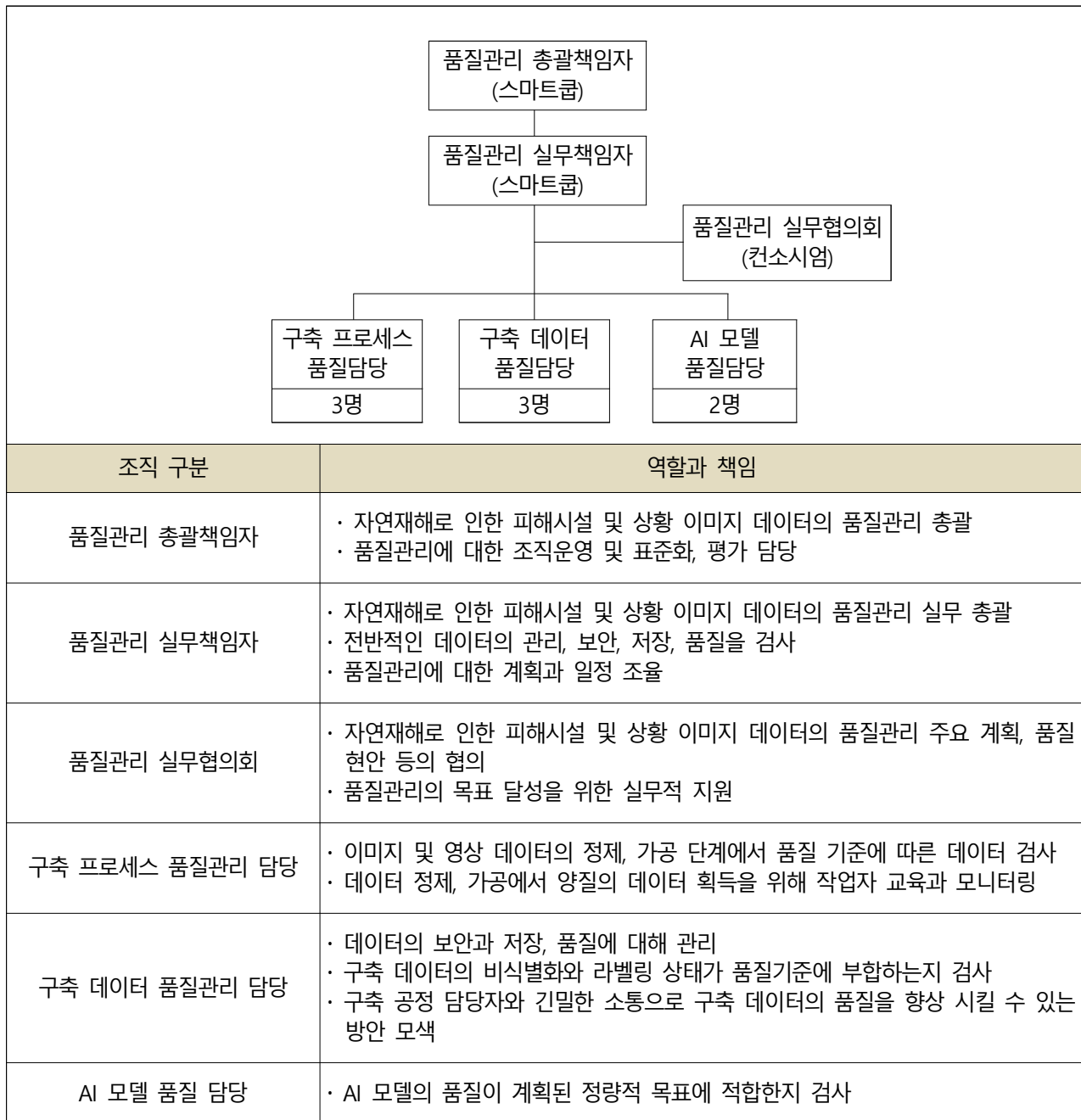
□ 정제 시 고려사항

획득/정제 시 고려사항	내용	예방 방안
중복성	1. 원시데이터 획득 시 장소나 구도가 같게 여러번 촬영 2. 동일한 영상을 사용하여 정제 및 편집을 반복하는 경우	1. 원시 데이터 획득 시 각 데이터를 모니터링하여 촬영자에게 피드백 2. 영상 정제시 작업자에게 영상 할당을 관리자가 직접 할당 (영상파일의 편집 및 복사 권한 분할을 통한 원천 통제)
표준화	정제 작업자의 개개인의 편차로 기준이 모호한 정제	정제 기준을 명확히 정하고 품질관리자의 지속적인 모니터링으로 기준 피드백
개인 정보 및 저작권	개인정보 및 사생활 정보 등 개인의 민감한 정보의 유출 가능성	개인정보에 대해 철저한 비식별화 혹은 데이터 획득에 대한 사전 동의서나 협약서등을 통한 관련서류 구비
데이터의 균형	데이터 획득 용이성이나 편의성에 따른 동일하거나 유사한 장소의 반복적 촬영	관리자를 통한 작업관리 및 명확한 임무전달

- 데이터 정제 작업 시 최종 데이터의 다양성을 위해 취득 환경, 재난, 상태 등의 환경 요소와 취득 대상 객체를 명확히 인지하여 필수 데이터가 폐기되지 않도록 유의하며 정제 작업 실시
- 데이터 정제 과정에 있어 정제 도구와 병행하여 육안 검수를 실시함으로써 최종 학습 데이터의 높은 품질수준을 확보 하도록 함
- 연속된 동일한 객체 및 장소의 데이터 처리를 통해 최종 학습 데이터의 다양성을 확보하도록 함
- 원활한 개인정보 비식별화 작업을 위해 정제 작업 중 사람 얼굴 및 자동차 번호판 등이 명확하게 식별 될 수 있도록 흐려짐, 초점 불일치, 화질 불량과 같은 특성을 갖는 데이터를 주의하여 정제 작업 실시
- 저작권 및 초상권 등에 침해되지 않게 해당 이미지를 제거하거나 인식이 불가능하도록 처리하며, 정보로서의 가치가 손상되지 않도록 주의

3.4. 수집·정제 조직

□ 품질관리 조직도



□ 품질관리 인력 구성

구분	성명	소속	역할	비고
총괄책임자	이역수	스마트쿵	품질관리 총괄	
실무책임자	최진욱	스마트쿵	전반적인 데이터 품질 관리, 각 관리자 역할과 일정 조율	- 겸임 불가

구분	성명	소속	역할	비고
실무협의회	한상훈	디노플러스	빅데이터 플랫폼 전문가	<ul style="list-style-type: none"> - 인공지능과 데이터 전문가로 구성된 자문단 - 개인정보, 초상권 등 관련 법률전문가 포함 고려
	이태현	디노플러스	AI 알고리즘 전문가	
	김지상	한구조엔지니어링	시설 전문가	
	김상식	한국건축구조기술사회	건축물 전문가	
	강성훈	인더스웰	정보화 시스템 전문가	
구축프로세스 품질관리	김영진	엠에스구조기술사사무소	구축프로세스 품질관리 총괄	
	김형욱	한구조엔지니어링	획득정제단계 품질관리 담당	
	남나경	인더스웰	라벨링공정 품질관리 담당	
구축데이터 품질관리	한순재	인더스웰	구축데이터 품질관리 총괄	
	전지혜	한국건축구조기술사회	원천,원시데이터 품질관리 담당	
	이우현	인더스웰	라벨링데이터 품질관리 담당	
AI 모델 품질관리	김남식	디노플러스	AI 모델 품질관리 책임자	
	조소은	디노플러스	AI 모델 품질관리 실무자	

□ 데이터 수집 도구

- ## □ 데이터 정제 도구

- ## - 메타데이터 파싱방법

센서	사용 도구	용도	비고
CAMERA	labelme tool 육안작업	영상->이미지 변환 중복/유사 데이터 처리	인지 용 데이터의 경우 유사 이미지 처리는 하지 않음
metadata	data parsing (using python)	중복/유사 데이터 처리	-

메타데이터 정보 정책					output	
<div> <div> <div>1</div> <div>2</div> <div>3</div> <div>4</div> <div>5</div> <div>6</div> <div>7</div> <div>8</div> <div>9</div> <div>10</div> <div>11</div> <div>12</div> </div> <div>정제</div> </div>	No	속성명	원본 유형	정책 유형	필수여부	단위(비치)
	1	image_id	촬영 기록 정보	number	필수	17,56655, 126,97801
	2	image_name	이치치 이름	string	비필수	-
	3	image_size	크기	string	비필수	-
	4	image_path	경로	string	비필수	-
	5	image_resolution	이치치 해상도	string	필수	FHD
	6	image_type	촬영 방식	number	필수	17,56655, 126,97801
	7	image_device	촬영 장치	string	필수	17,56655, 126,97801
	8	image_camera_name	카메라 이름	string	비필수	-
	9	image_lens	렌즈	string	비필수	-
	10	image_license	이치치 라이선스	string	필수	-
	11	image_type	데이터 형식	string	필수	JPG, PNG, TIFF
	12	image_size	파일 크기	string	비필수	KB
	13	image_length	영상 길이	number	비필수	-
	14	image_location	촬영 위치	string	비필수	-
	15	image_creator	녹화자	number	비필수	-
						수도권
						01_mobile 2020-04-12 15:24 FHD JPG 37.56655, 126.97801
						02_mobile 2020-04-12 15:35 FHD JPG 37.56654, 126.97800
						01_mobile 2020-04-12 21:05 FHD JPG 38.4821, 125.48124
						01_mobile 2020-04-12 21:11 FHD JPG 38.4821, 125.48124
						01_mobile 2020-04-12 21:26 FHD JPG 38.4821, 125.48124
						01_mobile 2020-04-12 21:27 FHD JPG 38.4821, 125.48124
						03_cctv 2020-05-08 07:44 FHD JPG 35.15557, 124.11587
						03_cctv 2020-05-08 07:44 FHD JPG 35.15557, 124.11587
						03_cctv 2020-05-08 07:44 FHD JPG 35.15557, 124.11587
						03_cctv 2020-05-08 07:44 FHD JPG 35.15557, 124.11587
						제주도

- 필드 오버로딩에 대한 검토
- 모순점이 발견된 데이터에 대해서 수정 변환 필요
- 수동으로 결측값 입력 및 결측값 데이터 개체 또는 속성의 제거
- 하나의 속성값이 없더라도 유사성을 계산하는데 미치는 영향이 크지 않다면 적용하여 제거 가능
- 데이터 간 결측값을 가진 속성들이 산재해 있다면 너무 많은 데이터는 제외

○ 비식별화

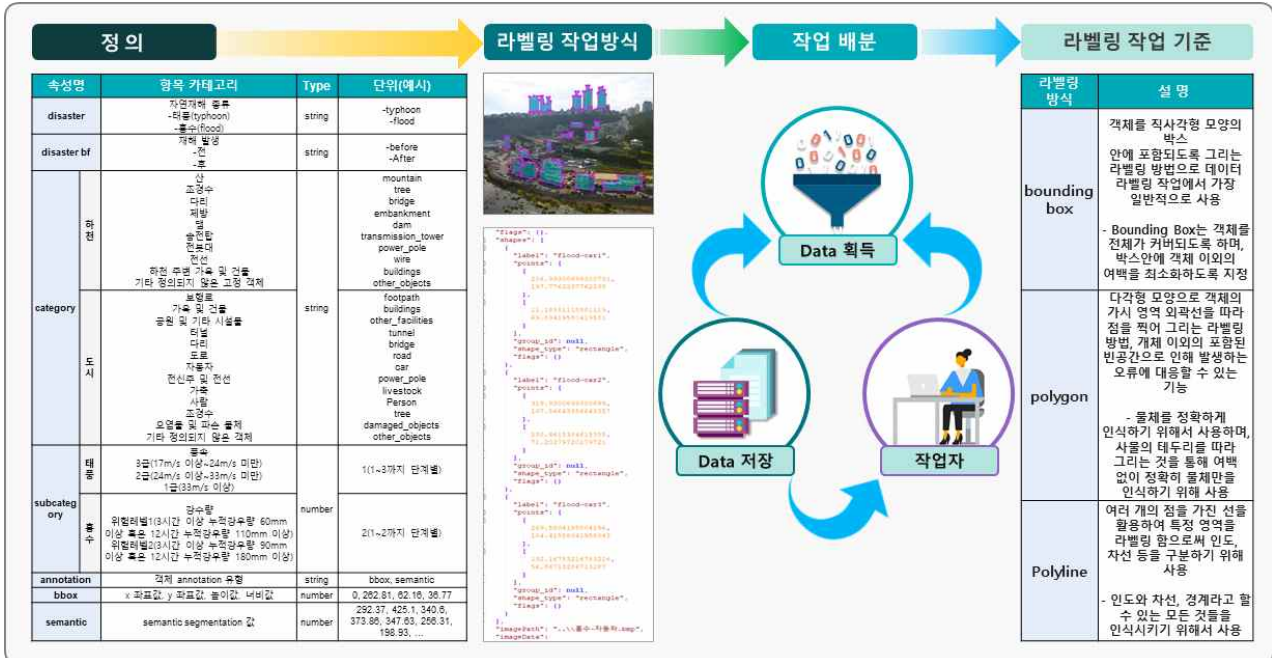
- 개인정보 보호를 위해 인물 얼굴, 자동차 번호판에 blur 효과를 주어 비식별화 처리
- 개인정보 비식별화 대상 및 예시

인물 얼굴	차량 번호판
 <p data-bbox="331 1120 550 1149">인물 얼굴 비식별화</p>	 <p data-bbox="954 1120 1216 1149">자동차 번호판 비식별화</p>

4. 데이터 가공

4.1. 라벨링 절차 및 방법

□ 자연 지해로 인한 피해 시설 이미지 데이터 어노테이션 프로세스



[어노테이션 프로세스]

□ 라벨링 및 어노테이션을 틀을 사용한 객체 정의 및 라벨링



[데이터 라벨링 객체 예시]

○ 저작 도구를 통해 이미지 어노테이션

○ 객체 인식을 위한 학습용 데이터 라벨링 분류 체계에 따라 정의된 객체를 Boundary Box, Polygon으로 데이터 라벨링 작업

□ 어노테이션 틀을 통한 라벨링 정보 JSON 파일 생성



[그림. JSON 파일 생성]

- 저작 도구를 통해 어노테이션 데이터 산출
 - 이미지 파일과 라벨링 데이터가 있는 json파일 생성
- 데이터의 DB저장 및 관리



[그림 18 어노테이션 데이터 저장]

- 이미지 파일과 라벨링 정보를 담고 있는 json 파일을 DB에 저장
- 데이터 구축 단계별로 데이터의 형태와 내용에 따라 저장경로와 구조 체계 정립

4.2. 라벨링 기준

□ 태풍 및 홍수로 인한 피해시설 이미지 데이터 라벨링 예시

○ 원시 데이터 및 정제 데이터에서 라벨링 정의 및 필요 데이터 선정 및 분류

[이미지 메타 데이터 예시]

No	속성명	항목 설명	Type	필수여부	단위(예시)
1	image[].mission	촬영 위치정보	number	필수	37.56655, 126.97801
2	image[].resolution	이미지 해상도	string	필수	FHD
3	image[].Date	촬영시간	number	필수	20210404_000001
4	image[].device	데이터 수집 도구	string	필수	01_mobile
5	image[].license	이미지 출처	string	필수	-
6	image[].type	데이터 형식	string	필수	JPG, JPEG, PNG
7	image[].location	홍수 지역 구분	string	선택	CF, RF
8	image[].section	수계 권역별 구분	number	선택	1~11
9	image[].stream	상류 하류 구분	string	선택	up, down
10	image[].rainfall_amount	강우량	number	선택	100

[Annotation 데이터 예시]

No	속성명		항목 카테고리	Type	필수 여부	단위(예시)
1	disaster		자연재해 종류 태풍(typhoon) 홍수(flood)	string	필수	-typhoon -flood
2	disaster bf		재해 발생 -전 -후	string	필수	before after
3	category	하천	제방(운동기구/놀이기구/기타) 댐 교량 및 교각	string	필수	levee dam bridge, pier
		도시	교통신호등 보행신호등 교통안전표지판(소형) 이정표표지판(대형) 사람 자동차 산사태 도로 건물			buildings car traffic light pedestrian traffic light car person traffic sign distance sign road

No	속성명		항목 카테고리	Type	필수 여부	단위(예시)
		공통	가로등 조경수 전신주 지하차도	string	필수	telegraph pole street_lamp tree underpass
		피해영역		string	필수	damaged area
		침수 피해영역		string	필수	flood damaged area
		부유물 피해영역		string	필수	flood debris damaged area
		토사 피해영역		string	필수	landslide damaged area
4	subcategory		위험레벨	number	필수	1(1~3까지 단계별)
5	annotation		객체 annotation 유형	string	필수	bbox, semantic
6	bbox		x 좌표값, y 좌표값, 높이값, 너비값	number	필수	0, 262.81, 62.16, 36.77
7	semantic		semantic segmentation 값	number	필수	292.37, 425.1, 340.6, 373.86, 347.63, 256.31, 198.93, ...

□ 위험 단계 카테고리에 대한 정의

○ 위험레벨 정의

- 인공지능 학습용 데이터셋 사업의 자연재해 데이터 구축 목적 및 내용에 따라 구축 데이터셋을 "자연재해"분야의 레벨을 정상(Lv.0), 재해(Lv.1)의 2개 레벨로 정의함

구분	레벨	레벨 정의 기준
정상	Lv.0	정상
재해	Lv.1	파손, 손상 및 복구불능

- 구체적인 취약레벨을 정의할 수 없는 객체에 대하여는 취약레벨과 재해레벨을 통합하여 제해레벨로 수집함

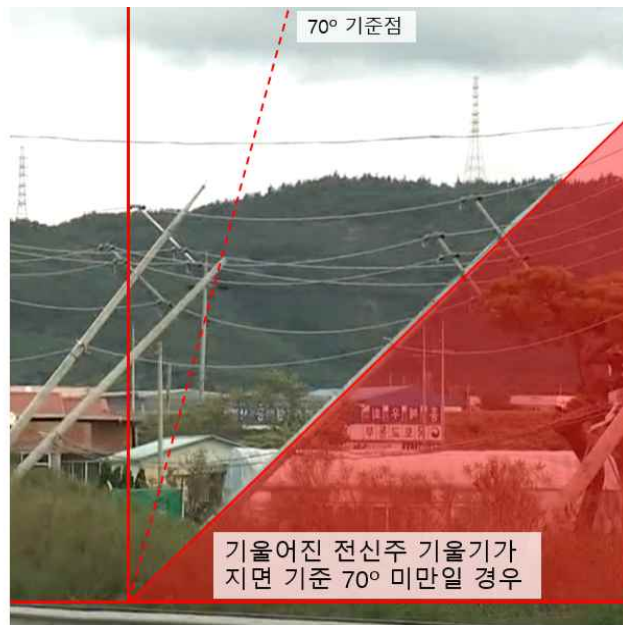
재해지역	객체	재해레벨	레벨정의	수량(건)
하천	제방 (운동기구/놀이 기구/기타)	정상 Lv.0	▪ 정상 상태의 제방	55,000
		재해 Lv.1	▪ 하천의 수위상승으로 인한 범람 및 붕괴 ▪ 제방의 하천 수위상승으로 인한 부유물 잔존	
	댐	정상 Lv.0	▪ 정상 상태의 댐	5,000

재해지역	객체	재해레벨	레벨정의	수량(건)
	교량 및 교각	재해 Lv.1	▪ 주의, 경계, 심각단계 및 그로 인한 부유물 생성	15,000
		정상 Lv.0	▪ 정상 상태의 교량 및 교각	
		재해 Lv.1	▪ 하천의 수위상승으로 인한 교각 및 교량의 침수 및 파손, 손상 ▪ 교각 및 교량에 수위상승으로 인한 부유물 잔존	
도시	교통신호등	정상 Lv.0	▪ 정상 상태의 교통신호등	15,000
		재해 Lv.1	▪ 교통신호등의 침수, 파손, 손상 및 침수로 인한 부유물 잔존	
	보행신호등	정상 Lv.0	▪ 정상 상태의 보행신호등	20,000
		재해 Lv.1	▪ 보행신호등의 침수, 파손, 손상 및 침수로 인한 부유물 잔존	
	교통안전표지판 (소형)	정상 Lv.0	▪ 정상 상태의 소형 교통안전표지판	15,000
		재해 Lv.1	▪ 소형 교통안전표지판의 침수, 파손, 손상 및 침수로 인한 부유물 잔존	
	이정표표지판 (대형)	정상 Lv.0	▪ 정상 상태의 대형 이정표표지판	15,000
		재해 Lv.1	▪ 대형 이정표표지판의 침수, 파손, 손상 및 침수로 인한 부유물 잔존	
	사람	정상 Lv.0	▪ 정상 상태의 사람	15,000
		재해 Lv.1	▪ 침수로 인해 사람의 발목이상 잠김	
	자동차	정상 Lv.0	▪ 정상 상태의 자동차	20,000
		재해 Lv.1	▪ 자동차 타이어 높이의 1/3이상 침수, 파손, 손상 및 침수로 인한 부유물 잔존	
	산사태	정상 Lv.0	▪ 정상 상태의 산능선, 봉우리 등	5,000
		재해 Lv.1	▪ 산사태 발생 및 산사태로 인한 토사 및 부유물 잔존	
	도로	정상 Lv.0	▪ 정상 상태의 도로(차선 구분 가능)	15,000
		재해 Lv.1	▪ 도로의 침수, 파손, 손상 및 침수로 인한 부유물 잔존	
	건물	정상 Lv.0	▪ 정상 상태의 건물	20,000
		재해 Lv.1	▪ 건물의 침수, 파손, 손상 및 침수로 인한 부유물 잔존	
공통 및 기타	가로등	정상 Lv.0	▪ 정상 상태의 가로등	25,000
		재해 Lv.1	▪ 가로등의 침수, 파손, 손상 및 침수로 인한 부유물 잔존	
	조경수	정상 Lv.0	▪ 정상 상태의 조경수	30,000
		재해 Lv.1	▪ 조경수의 침수, 파손, 손상 및 침수로 인한 부유물 잔존	
	전신주	정상 Lv.0	▪ 정상 상태의 전신주	25,000

재해지역	객체	재해레벨	레벨정의	수량(건)
		재해 Lv.1	▪ 전신주의 침수, 파손, 손상 및 침수로 인한 부유물 잔존	5,000
	지하차도	정상 Lv.0	▪ 정상 상태의 지하차도	
		재해 Lv.1	▪ 지하차도의 침수, 파손, 손상 및 침수로 인한 부유물 잔존	
총계(건)			300,000	

○ 객체 클래스별 위험레벨 기준 예시

- 전신주 레벨 1 기준 예시



- 가로등 레벨 1 기준 예시



○ 이미지 데이터 라벨링 예제

- 태풍 및 홍수로 인한 피해 발생시 객체 인식을 위한 학습용 데이터 라벨링 분류 체계
- 태풍 및 홍수 발생 전, 태풍 및 홍수 발생 후의 분류에 따른 라벨링

- 각각의 객체에 대한 데이터 형태, 라벨링 형태, 크기, 해상도 등 계획
- 라벨링 데이터 분류 체계

지역	code	상태	code	재난	객체	code
하천	01	정상(Lv.0)	01	하천	제방(운동기구/놀이기구/기타)	01
도시	02	재해(Lv.1)	02		댐	02
공통	03				교량 및 교각	05
				도시	건물	04
					산사태	06
					교통신호등	13
					보행신호등	14
					자동차	09
					사람	10
					교통안전표지판(소형)	15
					이정표표지판(대형)	16
				공통 및 기타	도로	17
					전신주	07
					가로등	08
					조경수	11
					지하차도	18
					파손 피해영역	12
					침수 피해영역	19
					부유물 피해영역	20
					토사 피해영역	21

- 이미지 데이터 라벨링 정보 세분화
- 객체분류에 따른 클래스 구조 정의
- 객체의 카테고리 정보
- 데이터 형태 라벨링
- 객체에 대한 픽셀단위 크기 분류
- 객체 해상도에 따른 분류
- 이미지 크기 및 해상도는 정확한 데이터 학습을 위하여 가능한 동일 스펙으로 구축 필요

4.3. 라벨링 규격

□ 어노테이션 구조

항목		타입	필수여부
한글명	영문명		
데이터셋명	info_name	string	Y
데이터셋상세설명	info_desc	string	Y
데이터셋 URL	info_url	string	Y
데이터셋 생성일자	info_data_created	string	Y
파일명	filename	string	Y
이미지 ID	image_id	int	Y
재해 종류	disaster	string	Y
어노테이션 ID	annotation_id	int	Y
세그멘테이션	segmentation	array	
경계 상자	bbox	array	
객체 카테고리	category	string	Y
객체 카테고리 ID	category_id	int	Y
제해 레벨	subcategory	int	Y

○ 어노테이션 구조 상세

- 객체 분류체계에 따른 어노테이션 정보 설계
- 포괄적 데이터 정보에 대한 유동적 설계
- 필수값 지정을 통한 어노테이션 누락 정보 방지
- 객체 분류체계에 따른 어노테이션 진행



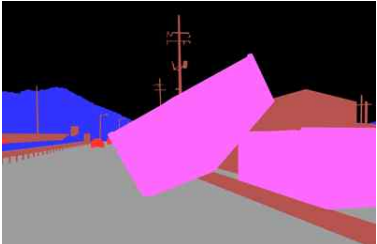
□ 라벨링 대상 설정

- 라벨링 분류 체계의 세분화를 통한 데이터 상세화
- 객체 데이터 라벨링 변경이 용이하도록 설계
- 태풍 및 홍수로 인한 피해 및 위험 데이터에서 식별해야 할 요소

지역	code	상태	code	재난	객체	code
하천	01	정상(Lv.0)	01	하천	제방(운동기구/놀이기구/기타)	01
도시	02	재해(Lv.1)	02		댐	02
공통	03				교량 및 교각	05
				도시	건물	04
					산사태	06

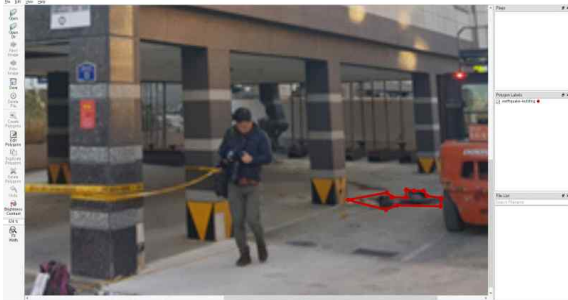

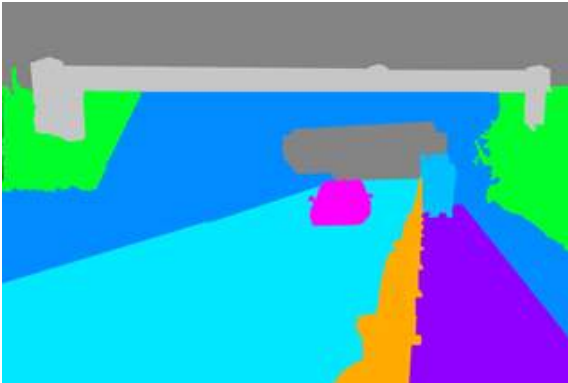
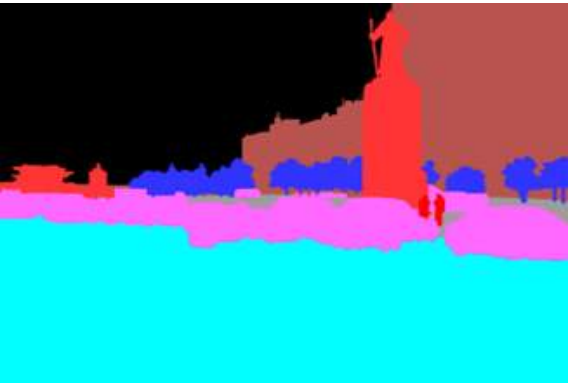




지역	code	상태	code	재난	객체	code
					교통신호등	13
					보행신호등	14
					자동차	09
					사람	10
					교통안전표지판(소형)	15
					이정표표지판(대형)	16
				공통 및 기타	도로	17
					전신주	07
					가로등	08
					조경수	11
					지하차도	18
					파손 피해영역	12
					침수 피해영역	19
					부유물 피해영역	20
					토사 피해영역	21

□ 라벨링 도구 정의

라벨링 방식	설명	이미지 예시
Bounding Box	<ul style="list-style-type: none"> 객체를 직사각형 모양의 박스 안에 포함되도록 그리는 라벨링 방법 객체 전체가 들어가도록 하되, 해당 객체 이외의 사물은 최소화하도록 할 것 	
Polygon	<ul style="list-style-type: none"> 다각형 모양으로 객체의 가시 영역 외곽선을 따라 점을 찍어 그리는 라벨링 방법 bounding box 방식 사용 시 여백이 많이 남는 객체에 활용 	
segmentation	<ul style="list-style-type: none"> polygon 방식을 사용하기에 재난 범위, 환경 등이 효율적이지 못 할 경우 사용 	

○ 식별 객체 별 사용 라벨링 예시

<p>① 조경수</p> 	<p>② 가옥 및 건물</p> 
<p>③ 간판</p> 	<p>④ 전신주</p> 
<p>⑥ 전봇대, 다리, 사람</p> 	<p>⑦ 다리, 제방</p> 
<p>⑧ 건물, 자동차, 산</p> 	<p>⑩ 전신주</p> 

<p>⑪ 건물</p> 	<p>⑫ 사람</p> 
<p>⑬ 자동차</p> 	<p>⑭ 차, 조경수</p> 
<p>⑮ 산, 댐</p> 	<p>⑯ 자동차</p> 
<p>⑰ 도로</p> 	<p>⑱ 제방</p> 

4.4. 가공 조직

□ 데이터 라벨링 조직도



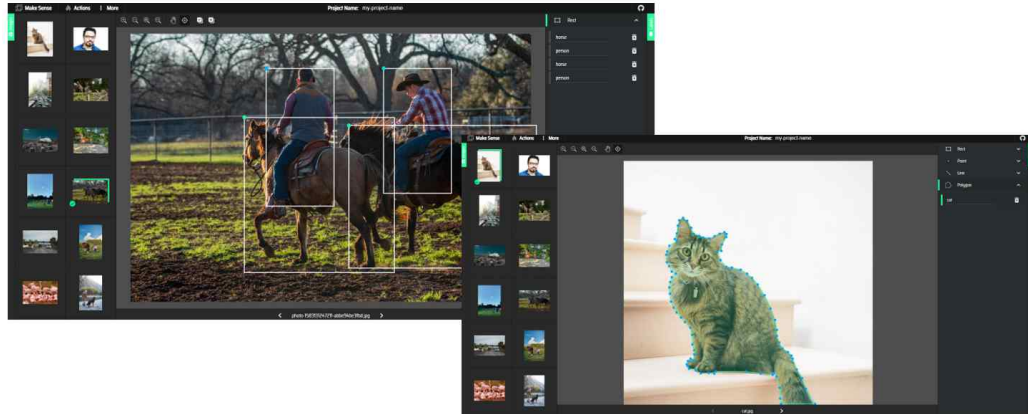
조직		역할 및 책임	교육 및 훈련
데이터 가공 책임자		<ul style="list-style-type: none"> - 데이터 가공 진척도 관리 - 데이터 중간 관리자 업무 관리 	<ul style="list-style-type: none"> - 인공지능 데이터셋 구축사업 교육 - 데이터 가공 작업 관리자 교육
데이터 가공 작업 관리자	데이터 레이블링 기준수립	<ul style="list-style-type: none"> - 레이블링 작업 기준 수립 - 라벨링 객체 추가 및 업데이트 	<ul style="list-style-type: none"> - 어노테이션 작업 교육 - 객체 정의 분류 안내
	데이터 가공작업 검수	<ul style="list-style-type: none"> - 레블링 작업 작업자 검수 안내 - 레이블링 작업 샘플링 검수 	<ul style="list-style-type: none"> - 가공 작업자 검수 교육
	데이터 가공작업량 관리	<ul style="list-style-type: none"> - 작업자 일일 작업 할당 관리 - 작업자 일일 프레임작업량 관리 	<ul style="list-style-type: none"> - 작업 업무 일지 작성 안내
데이터 가공 환경 관리자	데이터 구축 서버관리	<ul style="list-style-type: none"> - 산출물 구축량 관리 	<ul style="list-style-type: none"> - 이미지 다운로드 산출물 게재안내
	데이터 가공 서버관리	<ul style="list-style-type: none"> - 어노테이션 가공 서버 에러 관리 	<ul style="list-style-type: none"> - 데이터셋 추출 작업 방법 교육
데이터 가공인력 관리자	작업인력현황관리	<ul style="list-style-type: none"> - 출근 및 재택근무 현황 관리 - 근무 시간 관리 	<ul style="list-style-type: none"> - 출퇴근 제도 안내
	작업자 안전관리	<ul style="list-style-type: none"> - 체온 체크 및 방명록 관리 - 작업간 건강상태 관리 	<ul style="list-style-type: none"> - 성희롱 교육 - 작업 안전 수칙
	작업자 작업공간 관리	<ul style="list-style-type: none"> - 작업실 내 환기 	<ul style="list-style-type: none"> -

4.5. 라벨링 도구

□ 어노테이션 도구

○ 자체 저작도구 활용

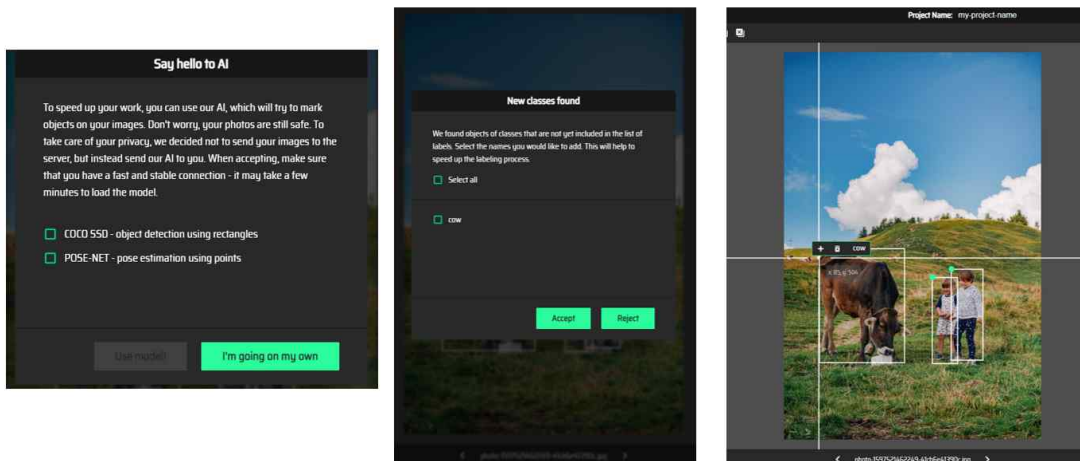
- 다양한 Segmantation과 자동 어노테이션 적용을 지원하는 어노테이션 도구 활용
- Bounding box, Polygon, semantic 등 다양한 형태의 라벨링 지원
- 서로 다른 속성을 가진 객체에 주석 입력 기능



[그림. 다양한 라벨링 형태 지원 어노테이션 도구]

○ 커스터마이징

- 기존 로컬 파일 기반 어노테이션 도구를 DB 및 웹서버 연동 가능하게 개발하여 활용
- 인공지능 모델링 파일을 적용하여 자동 어노테이션 기능이 있는 Make Sense 어노테이션 도구를 활용하여 어노테이션 작업 효율성 증대



[그림. 인공지능 모델링 활용 자동 객체 분류 기능 어노테이션 도구]

○ 수동 어노테이션 기능 사용

- 재해상황 및 피해상황에 대한 객체는 보편적인 형태의 객체와 차이점이 크기 때문에 자동어노테이션에 어려움이 있음
- 이에 따른 클라우드 워커 또는 라벨러가 어노테이션 도구로 Bounding Box, Polygon, Semantic Segmantation 형태의 라벨링을 직접 진행

5. 검수

5.1. 검수 절차 및 방법

□ 검사 절차 정의

- 다량의 데이터를 한정된 시간 내에 최적의 품질로 검사할 수 있도록 검사 단계 및 절차 수립
- 검사 프로세스는 학습용 데이터 구축 공정(획득, 정제, 라벨링) 각 단계별로 검사가 수행되는 형태를 기본으로 하며, 데이터 구축 공정 및 데이터 특성을 반영하여 적합한 절차를 수립

구분	내용
피해 및 위험 학습 데이터 획득	원시데이터는 태풍 및 홍수로 인한 피해 및 위험 데이터를 카테고리 분류하여 획득 및 활용
피해 및 위험 학습 데이터 정제	획득한 데이터에 대해 오류가 있다고 판단되는 이미지를 선정해 정제
피해 및 위험 학습 데이터 가공	객체 클래스 및 메타데이터 설계를 통해 객체 클래스를 분류하여 카테고리 별 피해 및 위험 데이터 가공
피해 및 위험 학습 데이터 검수	정제 및 가공 된 데이터를 대상으로 객체 누락 및 오분류 여부 결정
피해 및 위험 학습 데이터 확인	외부전문가 와 품질검수 전문가와의 검수 진행

□ 검사 규모

구분	데이터 구축량(단위 : 장)	제공 방식
정상 상태 이미지 데이터	180,000	원천 이미지(일부 영상 포함) + 라벨링 된 이미지 set
재해 상태 이미지 데이터	120,000	

- 피해 규모 판단에 필요한 다양한 객체 정보를 포함한 정상상태 18만 장, 재해상태 12만 장 이상의 이미지 데이터 검사

5.2. 검수 기준

□ 단계별 데이터 품질 검사 기준

- 품질 특성에 따른 태풍 및 홍수로 인한 피해 및 위험 AI 학습용 데이터 시험 지표를 기반으로 단계별 검증 기준 항목 선정

단계	구분	검사항목	검사 기준 및 내용
데이터 획득	획득 기준	해상도	▪ FHD(1920×1080) 및 FHD 공간 데이터
		Frame rate	▪ 30 FPS 이상
		파일 형식	▪ JPG, JPEG, PNG, MP4
		인코딩 일자	▪ 업로드 일자 1일 이내(클라우드소싱만 해당)
데이터 정제 (1차 편집)	데이터 구축 목적	장소의 중복성	▪ 촬영장소의 중복성 확인 ▪ 획득 단계에서 같은 장소, 같은 구도로 여러번 촬영
		객체 정보	▪ 객체 인식시 고려되어야 할 객체가 모두 검출되었는지 분석 ▪ 고려되어야 하는 객체 : 촬영 화면 내 모든 유형별 카테고리 라벨링 할 수 있도록 정확히 촬영되었는지 확인
	이미지 품질	흔들림	▪ 이미지 내 좌우 흔들림, 지속해서 비뚤어진 상태에서 촬영된 이미지 제거
데이터 정제 (2차 편집)	표준성	정제 작업자의 개개인의 편차로 기준이 모호한 객체	▪ 사람 얼굴
			▪ 자동차 번호판
			▪ 기타 민감정보(개인 전화번호, 보안시설 등)
	개인정보 및 저작권	이미지	▪ 데이터 획득단계에서 협약서, 동의서 등 관련서류 구비, 또는 철저한 비식별화 수행
데이터 가공	가공 기준	데이터의 균형	▪ 객체의 편향성 방지
		객체구분	▪ bounding, polygon, segmentation 등 객체의 구분
		객체 라벨링	▪ 잘못된 객체를 라벨링 한 경우 과검출, 라벨링 해야 할 객체를 빠뜨린 경우 미검출, 객체를 라벨링 하였으나 다른 객체로 인식하는 경우 오검출로 구분하여 검증 ▪ 충분히 멀리 떨어져 식별이 어렵거나 인공지능 모델에 영향을 주지 않는 객체 라벨링 대상에서 제외
		객체 선정	▪ 흐릿하거나 인공지능 모델에 영향을 주지 않는 객체는 배제

□ 수행단계별 검사 항목

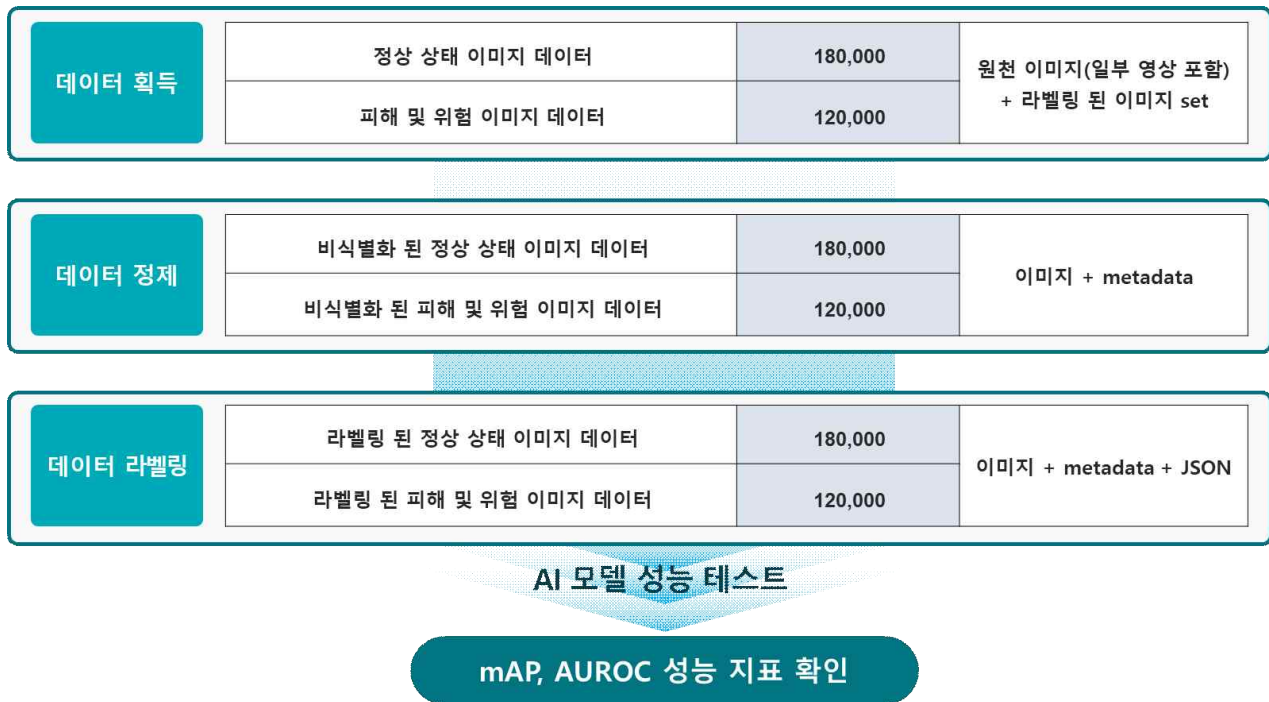
○ 검사 공정별 주요 검사 항목

- 구축 공정(획득, 정제, 라벨링) 별로 공통적으로 적용할 수 있는 검사 요구사항을 고려하여 검사 항목정의
- 검사 항목은 데이터 및 절차 측면에서 적합성·정확성·유효성, 준비성·완전성·유용성 지표를 측정 가능

검사 절차	영상(동적/정적) 이미지 공통 항목	요구사항
1차 검사 (획득)	법·제도 준수	원시데이터 획득 시 관련 법·제도적 규정 등을 반드시 준수하여야 함
	사실적인 획득 환경 구성	원시데이터를 인위적인 환경과 조건 하에 획득해야 하는 경우 사실적인 획득 환경을 구성하여야함
	장소의 중복성	획득 단계에서 같은 장소, 같은 구도로 여러번 촬영 필요
	편향성 방지	데이터 편향을 방지하기 위한 절차를 마련하여야함
2차 검사 (정제)	정제 기준의 명확성	데이터 사용 목적에 적합한 정제 기준 수립 여부 검사
	중복성 방지	데이터 정제 후 정보 비교 후 중복도 여부 검사
	개인정보	데이터 획득단계에서 협약서, 동의서 등, 관련서류 구비, 또는 철저한 비식별화 진행
	데이터의 균형	객체의 편향성 방지
	정제 작업 방식	데이터 특성 및 활용 목적에 맞는 적절한 정제 방식 선정 여부 및 선정 기준 타당성 여부 검사
3차 검사 (라벨링)	객체구분	bounding, polygon, segmentation 등 객체의 구분
	객체 라벨링	잘못된 객체를 라벨링 한 경우 과검출, 라벨링 해야 할 객체를 빠뜨린 경우 미검출, 객체를 라벨링 하였으나 다른 객체로 인식하는 경우 오검출로 구분하여 검증, 충분히 멀리 떨어져 식별이 어렵거나 인공지능 모델에 영향을 주지 않는 객체 라벨링 대상에서 제외
	객체 선정	흐릿하거나 인공지능 모델에 영향을 주지 않는 객체는 배제
4차 검사 (전수)	부적합 판정 데이터 분포 확인	데이터의 오류율, 특성 분포 확인을 통한 데이터 수집, 정제, 라벨링, 부문 최적화
	외부 검사자	외부 검사자(TTA 등), 품질관리 점검자와의 협업을 통한 전수검사

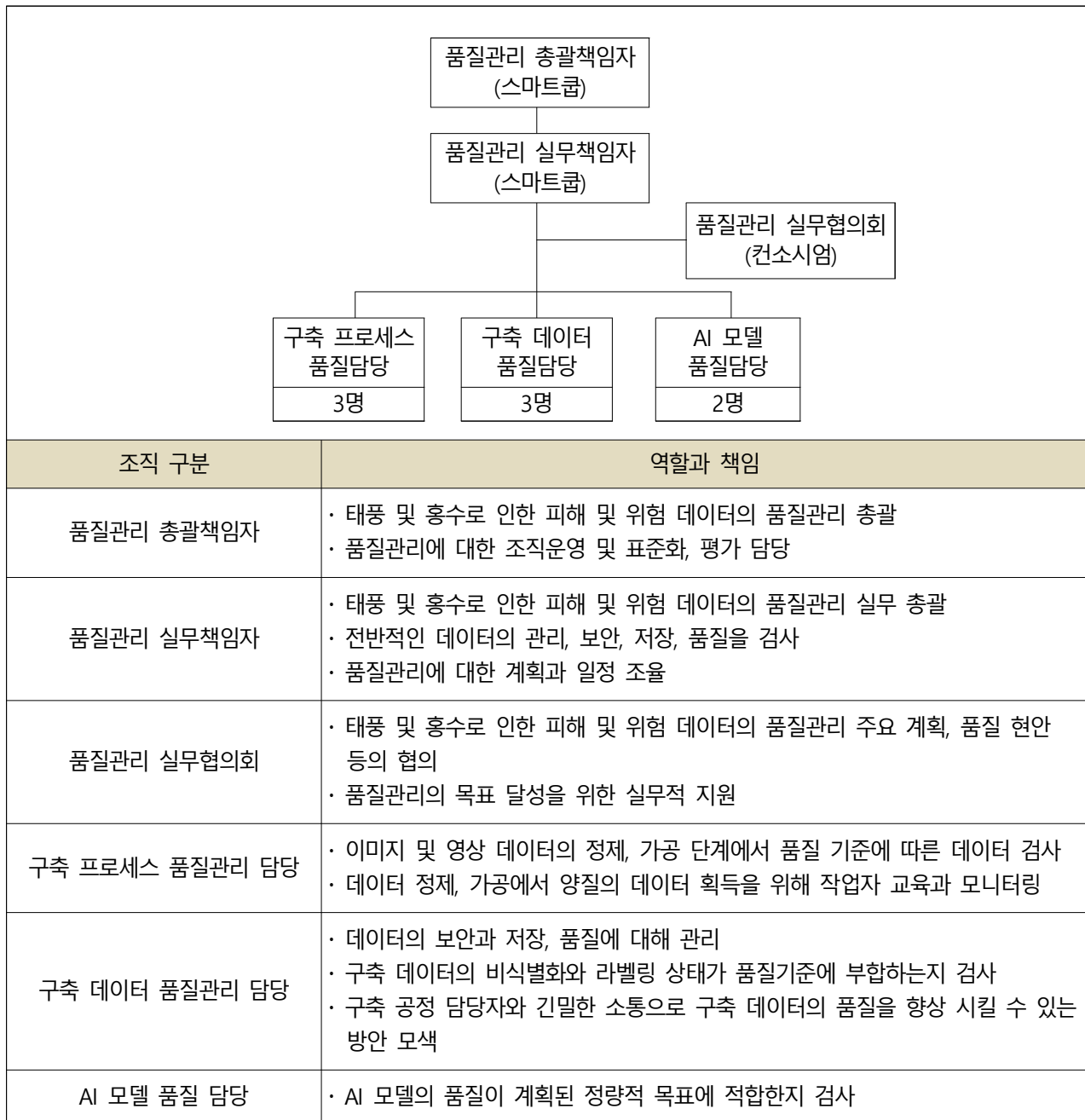
□ 1cycle 중심 검사 방법

○ 1cycle 세부 plan



5.3. 검수 조직

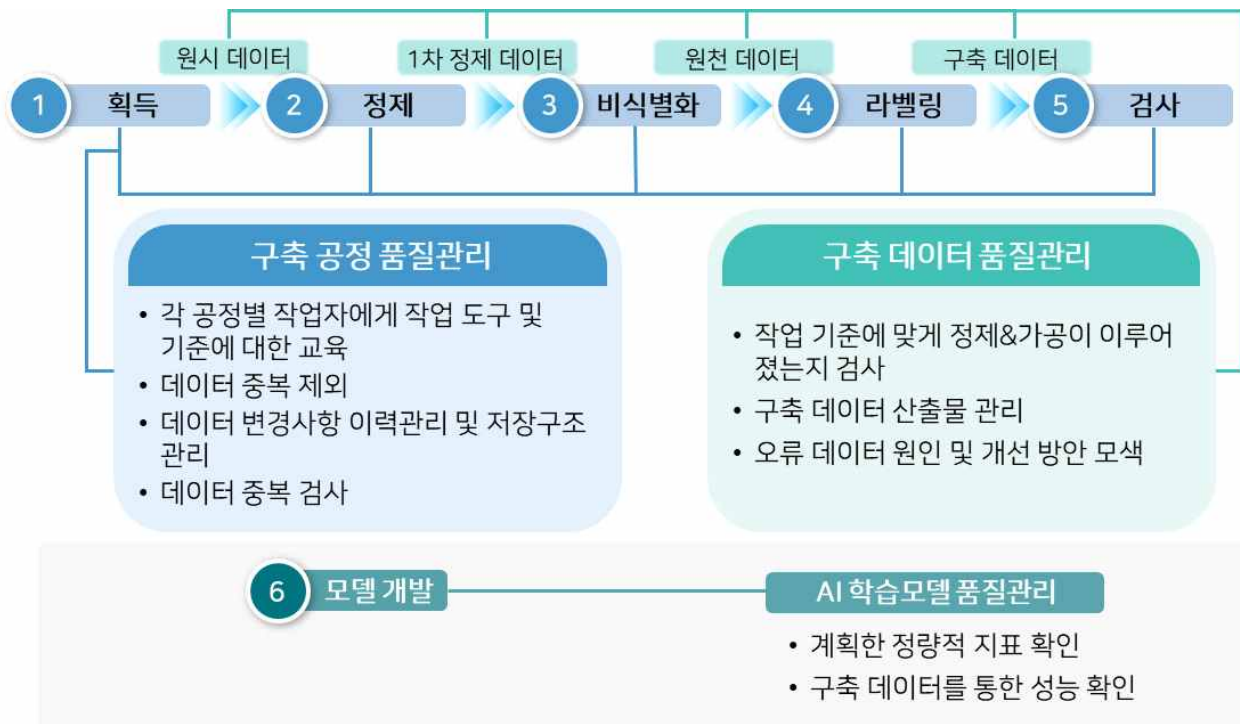
□ 품질관리 조직도



□ 품질관리 인력 구성

구분	성명	소속	역할	비고
총괄책임자	이역수	스마트쿵	전체 컨소시엄 관리 및 총괄	
실무책임자	최진욱	스마트쿵	전반적인 데이터 품질 관리, 각 관리자 역할과 일정 조율	- 겸임 불가

구분	성명	소속	역할	비고
실무협의회	한상훈	디노플러스	빅데이터 플랫폼 전문가	- 인공지능과 데이터 전문가로 구성된 자문단 - 개인정보, 초상권 등 관련 법률전문가 포함 고려
	이태현	디노플러스	AI 알고리즘 전문가	
	지상준	엘씨씨코리아	수질 전문가	
	강성훈	인더스웰	정보화 시스템 전문가	
구축프로세스 품질관리	이춘경	엘씨씨코리아	구축프로세스 품질관리 총괄	
	남나경	인더스웰	획득정제단계 품질관리 담당	
	한순재	인더스웰	라벨링공정 품질관리 담당	
구축데이터 품질관리	권보민	엘씨씨코리아	구축프로세스 품질관리 총괄	
	강국현	엘씨씨코리아	원천,원시데이터 품질관리 담당	
	전병훈	인더스웰	라벨링데이터 품질관리 담당	
AI 모델 품질관리	김남식	디노플러스	AI 모델 품질관리 책임자	
	조소은	디노플러스	AI 모델 품질관리 실무자	



[품질관리 조직구성 및 역할]

□ 품질관리 교육 계획


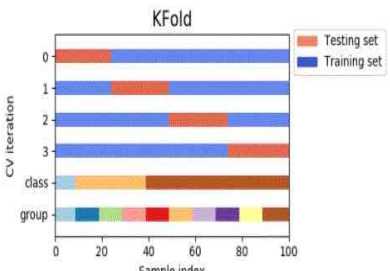
회차	교육과정	교육내용	교육일정	교육대상	비고
1회차	품질관리 기본교육	인공지능 학습용 데이터 품질관리 가이드 작성 방법 인공지능 학습용 데이터 구축계획 작성 방법	사업착수 후 14일 이내	구축사업에 참여하는 수행기관 및 참여기관 전원	필수
2회차	원시데이터 품질관리 실무교육	- 인공지능 데이터 구축 사업소개 - 원시데이터 획득 방법 교육 - 원시데이터 품질 기준 교육 - 1차 정제 기준 방법 교육	사업착수 후 수시교육	원시데이터 구축 실무자(클라우드 워커 포함)	한국건축구 조기술사회
3회차	원천데이터 품질관리 실무교육	- 인공지능 데이터 구축 사업소개 - 원천데이터 품질 기준 교육 - 원천데이터 품질 검사 방법 교육 - 비식별화 도구 사용방법 (오류데이터 탐색 및 수정)	사업착수 후 수시교육	비식별화 실무자 (클라우드워커 포함)	한국건축구 조기술사회
4회차	어노테이션 품질관리 실무교육	- 인공지능 데이터 구축 사업소개 - 라벨링 품질 기준 교육 - 라벨링 품질 검사 방법 교육 - 어노테이션 도구 사용방법 (오류데이터 탐색 및 수정)	사업착수 후 수시교육	데이터 가공 실무자(클라우드 워커 포함)	인더스웰
5회차	최종 구축데이터 품질검사 실무교육	- 인공지능 데이터 구축 사업소개 - 구축데이터 품질 기준 교육 - 구축데이터 품질 검사 방법 교육 - 저작도구 사용방법 (오류데이터 탐색 및 수정)	사업착수 후 수시교육	구축 데이터 품질 실무자 (클라우드워커 포함)	디노플러스

5.4. 검수 도구

□ 품질관리 도구

품질검사 영역	구문 정확성 검사	의미 정확성 검사	학습모델 유효성 검사
도구명	parsing 도구	정확성 검사 도구	AI 모델 유효성 검사 도구
설명	python 기반의 parsing 도구를 사용하여 구문 정확성 검사	이미지 데이터 품질 검사	학습용 이미지 데이터 유효성 검사
도구유형	상용SW	자체 개발	자체 개발
주요기능	<ul style="list-style-type: none"> - 검사대상 데이터 등록관리 - 검사기준관리 - 품질검사관리 - 품질검사이력관리 - 보고서 	<ul style="list-style-type: none"> - 검사대상 데이터 등록관리 - 검사기준 관리 - 품질검사관리 - 품질검사 이력관리 - 보고서 	<ul style="list-style-type: none"> - 검사 기준 관리 - 품질검사 관리 - 품질 검사 이력관리 - 보고서
사용환경	<ul style="list-style-type: none"> - OS : Windows 10, 64bit - SW : python application 	<ul style="list-style-type: none"> - OS: Windows 10 - S/W: Python 3.6 이상 	<ul style="list-style-type: none"> - OS : Windows 10 - S/W: Python 3.6 이상

□ 품질관리 도구 상세

구문 정확성 검사	의미 정확성 검사	학습모델 유효성 검사
<p>총 개수 : 1445</p> <p>목적차량(특장차): 기타특장차의 개수는 68개, 비율은 4.71% 이다</p> <p>보행자: 성인(노인포함)의 개수는 192개, 비율은 13.29% 이다</p> <p>보행자: 어린이의 개수는 7개, 비율은 0.48% 이다</p> <p>보행자: 자전거의 개수는 18개, 비율은 0.69% 이다</p> <p>이륜차: 오토바이의 개수는 14개, 비율은 0.97% 이다</p> <p>일반차량: SUV/승합차의 개수는 458개, 비율은 31.78% 이다</p> <p>일반차량: 버스(소형,대형)의 개수는 175개, 비율은 12.11% 이다</p> <p>일반차량: 세단의 개수는 419개, 비율은 29.00% 이다</p> <p>일반차량: 통학버스(소형,대형)의 개수는 31개, 비율은 2.15% 이다</p> <p>일반차량: 트럭의 개수는 71개, 비율은 4.91% 이다</p>		

5.5. 기타 품질관리 활동

□ 품질관리 기준

품질관리 영역	품질지표	세부지표		적용 여부	요구사항ID
프로세스 품질	준비성	계획 수립	절차	적용	RP-009, RP-011, RP-012
			조직	적용	RP-001
			도구	적용	RP-011
			위험관리	N/A	-
		체계 준수	보안	적용	RP-004, RP-005
			법제도	적용	RP-005
	완전성	획득		적용	RD-001, RD-002, RD-004, RD-005, RD-007
		정제		적용	RP-002, RP-006, RP-007, RP-008
		라벨링		적용	RP-012
	유용성	사용편의		적용	RD-007
		유연성		적용	RP-003
데이터 품질	적합성	기준 적합성	다양성	적용	RP-006
			신뢰성	적용	RP-007, RP-008, RD-006, RD-008
			충분성	적용	-
			균일성	적용	RP-007, RP-010, RD-002
			사실성	적용	RD-006
			공평성	적용	RD-001, RD-002
		기술 적합성	파일포맷 준수율	적용	RD-001, RD-007
			해상도 준수율	적용	RD-002
		영상 품질 준수율		적용	RD-002, RD-003
		통계적 다양성	클래스분포	적용	RM-001
			인스턴스분포	적용	-
	정확성	의미 정확성	정확도	적용	RP-010, RD-008
		구문 정확성	데이터구조 오류	적용	RD-008, RD-009, RD-010
			입력값 오류	적용	
			데이터 형식 오류	적용	
학습모델 품질	유효성	모델 정확도		적용	RM-001
		객체 세그멘테이션		적용	RM-002

□ 품질검사 절차

○ 프로세스 품질검사

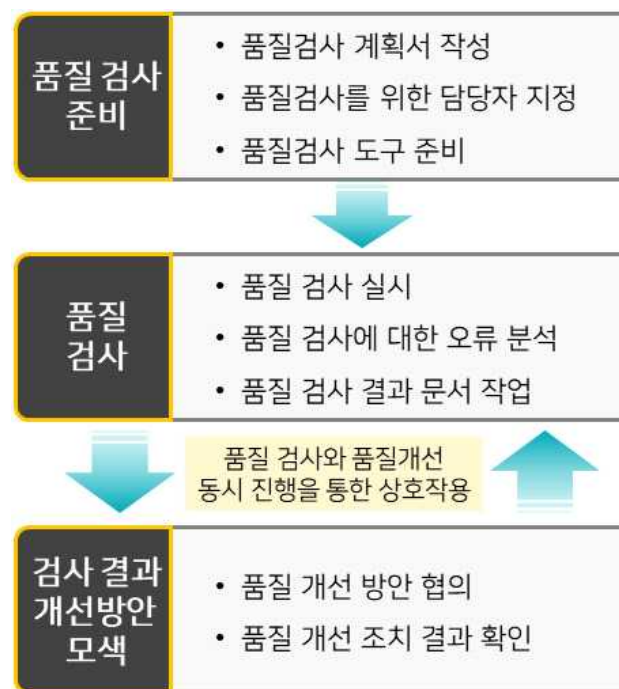
단계	세부 업무 내용	일정		담당자	주요 산출물
		시작일	종료일		
1. 품질검사준비	품질검사 계획서 작성	'21.07.01	'21.07.31	최진욱	- 품질검사계획서
	품질검사 담당자 배정	'21.07.01	'21.07.31		
	품질검사 도구 준비	'21.08.01	'21.08.31		
2. 품질검사실시	품질검사 실시	'21.09.01	'22.12.31	최진욱	- 품질검사결과서
	품질오류 분석	'21.09.01	'22.12.31		
	결과문서 작업	'22.01.31	'21.02.28		
3. 품질개선조치	품질 개선 방안협의	'21.08.01	'21.12.31	최진욱	- 품질개선결과서
	개선 조치 결과 확인	'21.08.01	'21.12.31		

○ 데이터 품질검사

단계	세부 업무 내용	일정		담당자	주요 산출물
		시작일	종료일		
1. 품질검사준비	품질검사 계획서 작성	'21.07.01	'21.07.31	최진욱	- 품질검사계획서
	품질검사 담당자 배정	'21.07.01	'21.07.31		
	품질검사 도구 준비	'21.08.01	'21.08.31		
2. 품질검사실시	품질검사 실시	'21.09.01	'22.12.31	최진욱	- 품질검사결과서
	품질오류 분석	'21.09.01	'22.12.31		
	결과문서 작업	'22.01.01	'22.02.28		
3. 품질개선조치	품질 개선 방안협의	'21.09.01	'21.12.31	최진욱	- 품질개선결과서
	개선 조치 결과 확인	'21.09.01	'21.12.31		

□ 학습모델 품질검사

단계	세부 업무 내용	일정		담당자	주요 산출물
		시작일	종료일		
1. 품질검사준비	검사 일정 협의, 품질 계획서 작성, 품질검사 기준 확정, 품질 검사 범위, 품질검사방법	'21.07.01	'21.07.31	최진욱	- 품질검사계획서
2. 품질검사실시	검사결과확인, 검사결과통보, 결과문서작업	'21.09.01	'21.12.31	최진욱	- 품질검사결과서
3. 품질검사조치	품질오류원인분석, 개선방안협의, 개선조치이행, 개선조치결과방안, 개선조치결과확인, 결과문서 작업	'22.01.01	'22.2.28	최진욱	- 품질개선결과서



[품질검사 절차 관계도]