인공지능 데이터 구축·활용 가이드라인
– 도로장애물,표면 인지 영상(광역시, 고속도로, 국도 등)–

인공지능 데이터 구축	사업 총괄	
	데이터 설계	
	원천데이터 수집 및 정제	
	데이터 가공	
	데이터 검수	
	크라우드 소싱	
	저작도구 개발	
	AI모델 개발	
	응용 서비스 개발	
가이드라인 작성	건국대학교 산학협력단	김상권
	건국대학교 산학협력단	정지민
가이드라인 버전	Ver 1.1 2021. 02. 06	

목 차

1. 데이터 명세 정보

1.1 데이터 정보 요약

데이터 이름	도로장애물/표면 인지 영상 (광역시, 고속도로, 국도 등)	
활용 분야	도로 유지보수, 자율주행 회피 거동	
데이터 요약	자율주행 중 도로상의 장애물 및 도로 표면의 이상 상태를 영상기반으로 인식할 수 있는 인공지능 개발을 위한 학습 데이터	
데이터 출처	자율주행	
데이터 이력	배포버전	1.0
	개정이력	신규
	작성자/ 배포자	건국대학교 산학협력단/ 건국대학교 산학협력단

1.2 데이터 포맷

	예시	데이터 항목	데이터 형식
영상		원천 데이터	*.mp4
이미지		객체 영역 객체 종류 노면 상태	*.png *.json "name": "0a0a0b1a-7c39d841.jpg", "attributes": { "weather": "clear", "scene": "highway", "timeofday": "daytime" > },

			"timestamp": 10000,
			"labels": [
			{
			"category": "car",
			"attributes": {
			"occluded": true,
			"truncated": false,
			"trafficLightColor": "none"
			},
			"manualShape": true,
			"manualAttributes": true,
			"box2d": {
			"x1": 555.647397,
			"y1": 304.228432,
			"x2": 574.015906,
			"y2": 316.474104
			},
			"id": 109344
			},

GNSS	$GPRM ,083559.00,A,4717.11437,N,00 833.91522,E,0.004,091202	시간 상태 위도 North/South 경도 East/West 속도 날짜	

1.3 어노테이션 포맷

No		항목명	항목설명	타입	필수구분	단위
1		데이터셋정보	데이터셋 전체에 관한 전반적인 정보를 포함하는 메타데이터	Object		
1	1-1	데이터셋명		String	Y	
	1-2	데이터셋상세설명		String		
	1-3	데이터셋URL		String		
	1-4	데이터셋생성일자		String	Y	
2		이미지정보	데이터셋을 구성하는 각 이미지에 대한 메타데이터 및 학습 데이터	List		
2	2-1	이미지식별자		String	Y	
	2-2	이미지너비		Number	Y	pixel
	2-3	이미지높이		Number	Y	pixel
	2-4	이미지파일명		String	Y	
	2-5	이미지라이선스		String		
	2-6	이미지촬영일자		String		
	2-7	이미지촬영날씨		String	Y	
	2-8	이미지촬영시간대		String	Y	
	2-9	원본영상정보		String	Y	
	2-10	프레임순서		String	Y	
3		어노테이션정보	데이터셋의 어노테이션에 대한 메타데이터 및 학습 데이터	List		
3	3-1	어노테이션식별자		String	Y	
	3-2	연관이미지식별자		String	Y	
	3-3	어노테이션속성		Object		
	3-4	어노테이션 바운딩박스		List		
	3-5	어노테이션 이미지		Image		

1.4 데이터 구성

1.4.1 데이터 Naming 규칙

- <차량구분>_<영상장치>_<동영상 No.>_<촬영일자>_<비식별화>_<카메라No>_<촬영지역>_<날씨구분>_<도로상태>_<촬영시간구분>_<PNG No>.PNG
  - 예) 2020년 10월30일 11시 30분, 부산에서 구름 낀 날씨에 촬영한 이미지
    V1F_HY_0002_20201030_113045_E_CHO_Busan_Cloud_Mainroad_Day_0005.png

구분	이름	설명
차량구분	V0F, V1F, V2F, ... VnF	차량 및 운전자 고유번호
영상장치	HY_0002, HY_0015, ...	동영상(1분, MP4) 고유 No.
촬영일시	촬영일 : YYYY/MM/DD	(년월일) _20201118_
	촬영시간 : hh/mm/ss	시/분/초 _131958_
비식별화	N / E	N : 비식별 X E : 비식별 O
카메라(채널)	CH0, CH1, ...	카메라 위치, 채널 No.
촬영 지역	Seoul / Busan	수도권, 광역시 구분
날씨 구분	Sun / Cloud / Rain / Fog / Snow	맑음/흐림/비/안개/눈 등
도로 상태	Frontback / Highway / Kidzone / Mainroad / Industrialroads	도심(골목길), 고속도로 어린이보호구역), 국(지방)도, 항만/공단
촬영시간 구분	Day / Night / Sunrise / Sunset	낮/밤/일출/일몰 등
PNG No.	_0005	이미지 생성시 PNG 번호

1.5 데이터 통계

1.5.1 데이터 구축 규모

1.5.1.1 영상 및 이미지 데이터의 규모 표현

- 도로 촬영 영상 300시간
- 연속촬영영상(sequence) 형태로 구성된 도로 촬영 이미지 100만 장
  - 약 100만장의 annotation 수행
  - 유효영상 : 대상 객체가 포함된 1분 내외의 영상 파일 300시간 이상
  - 유효이미지 : 위 유효영상에서 PNG 추출한 이미지파일(PNG) 100만장
  - 예측불가 장애물의 경우 가상의 환경을 만들어 촬영하므로 같은 장소에서 촬영 가능

1.5.1.2 GNSS 데이터의 규모 표현

- 도로 촬영 이미지와 동기화된 절대 위치정보 100만 건

1.5.2 데이터 분포

1.5.2.1 객체별 분포

이미지 프레임				AI 학습용 데이터 구축량
객체 바운딩박스	동적 객체	예측불가 동적객체	도로상에 출현하는 고라니, 사슴 보행자	5만
	정적 객체	예측불가 정적객체	화물차에서의 낙하물(상자), 라바콘, 공사표지판, 쓰레기	30만
		도로위 낙석	산사태 등의 암석	5만
	노면	포트홀	포트홀	10만
		보수완료 포트홀	정상도로에 보수완료된 포트홀	20만
		맨홀	정상도로에 맨홀	
객체 시멘틱 세그맨테이션	노면	크랙	크랙	30만

1.5.2.2 시간대 및 날씨 별 분포

- 맑음, 흐림, 안개 또는 비에 대해 각각 60%, 30%, 10% 비율로 수집
  - 기상조건이 충족되지 않아 불가피하게 데이터 수집이 어려울 경우에는 충족 가능한 날씨 환경에서 최대한 편향되지 않은 데이터 취득

1.5.3 기타 활용 통계

Table 1
Method performance for the test of 70 pavement images.

Performance	
Total TP	31
Total FP	7
Total TN	42
Total FN	5
Accuracy	85.9%
Precision	81.6%
Recall	86.1%

(출처 : Koch, Christian, and Ioannis Brilakis. "Pothole detection in asphalt pavement images." Advanced Engineering Informatics 25.3 (2011): 507-515.)

TABLE I
GROUND TRUTH PARAMETERS

Class	24.03.2011	19.03.2010
Large potholes	3	3
Small potholes	18	18
Pothole clusters	30	30
Gaps	40	25
Drain pits	17	29
Total	**108**	**105**

TABLE II
ACCELEROMETER DIFFERENCES BETWEEN ANDROID SMART-PHONES, AVERAGED OVER 10 MINUTE DRIVE

Device	Sampling rate (Hz)	Z-axis StdDev (g)
Samsung i5700	26	0.3076
Samsung Galaxy S	98	0.1171
HTC Desire	52	0.1215
HTC HD2	47	0.1242

TABLE III
TRUE POSITIVE RATE OF THE FOUR USED ALGORITHMS

Class	Z-THRESH	Z-DIFF	STDEV(Z)	G-ZERO
Large potholes	3 (100%)	3 (100%)	3 (100%)	3 (100%)
Small potholes	15 (83%)	16 (89%)	16 (89%)	14 (78%)
Pothole clusters	25 (83%)	27 (90%)	27 (90%)	27 (90%)
Gaps	31 (78%)	36 (90%)	30 (75%)	27 (68%)
Drain pits	10 (59%)	17 (100%)	11 (65%)	8 (47%)
Total	**84 (78%)**	**99 (92%)**	**87 (81%)**	**79 (73%)**

(출처 : Mednis, Artis, et al. "Real time pothole detection using android smartphones with accelerometers." 2011 International conference on distributed computing in sensor systems and workshops (DCOSS). IEEE, 2011.)

Performances	The existing method	The proposed method
Total TP	22	44
Total FP	18	11
Total TN	24	31
Total FN	38	16
Accuracy	45.1%	73.5%
Precision	55.0%	80.0%
Recall	36.7 %	73.3%

(출처 : Ryu, Seung-Ki, Taehyeong Kim, and Young-Ro Kim. "Image-based pothole detection system for ITS service and road management system." Mathematical Problems in Engineering 2015 (2015).)

1.6 원시데이터 특성

1.6.1 대상분류

- 복합

1.6.2 제약조건

- 일부 제약있음
  - 도로 주행 영상 수집 시 특정 객체가 존재하는 경우
  - 특정객체: 낙하물, 포트홀, 크랙 등

1.6.3 속성

- 해상도
  - 1920*1080 혹은 1280*720
- 초당 프레임수
  - 30 프레임 이내

1.7 기타정보

1.7.1 포괄성

데이터 분류	경로 및 지역 종류	데이터 수집 비율
수도권 지역 데이터	고속도로	5%
	국도(지방도로 포함)	35%
	도심(골목길 포함)	30%
	항만/공단	30%

1.7.2 독립성

- 사람 얼굴, 간판, 번호판 등 사생활 침해의 소지가 있는 민감 정보 정제 필요

1.7.3 유의사항

- 자율주행 시스템 적용을 통한 관련 기업들의 기술력 제고 및 AI 산업 관련 국가경쟁력 강화
- 데이터·AI 혁신기술 중심으로 산업구조 전환되는 경제적 파급효과
- 공공 데이터 제공을 통한 국민의 다양한 활용 및 사회 전반의 참여기회 제공관련 연구

2. 데이터 구축 가이드

2.1 데이터 구축 개요

2.1.1 데이터 수집

2.1.1.1 취득, 저장, 동기화 검수

2.1.2 데이터 정제

2.1.2.1 사용불가 데이터 분류

2.1.2.2 개인정보 비식별화

2.1.3 데이터 가공

2.1.3.1 단계별 어노테이션

2.1.4 데이터 검수

2.1.4.1 어노테이션 검사

2.1.4.2 유효성 검사

2.2 문제정의

2.2.1 임무 정의

2.2.1.1 자율주행 자동차가 주행 중 입는 손상을 최소화하고, 주행 가능 여부, 회피 여부 등 도로 상태 판단을 위한 학습 데이터 구축

2.2.2 데이터 구축 유의사항

2.2.2.1 개인정보 비식별화

- 수집된 데이터에 대한 개인 초상권 및 차량번호 등 개인정보 비식별화
- 얼굴 정보 비식별화
- 차량 번호판 및 간판(상표 등) 비식별화

2.3 획득·정제

2.3.1 원시데이터 선정

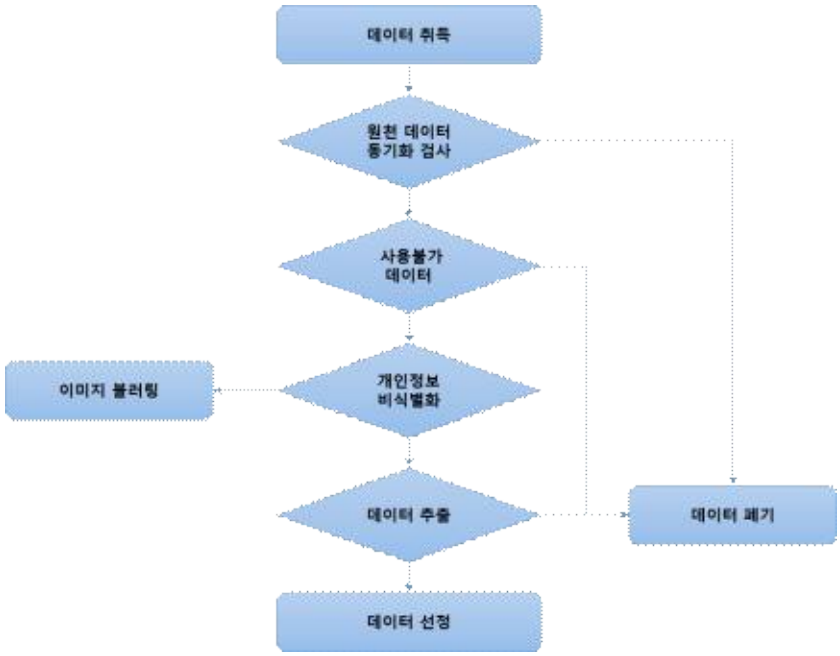
2.3.1.1 원본 영상 이미지

- 도로 내 다양한 형태의 장애물, 포트홀, 크랙등을 인식
- 원천 영상데이터을 활용하여 도로 내 포트홀 보수 혹은 보수를 위한 데이터 통보와 도로 주행시 안전 운행을 위한 DataBase로 활용 가능

2.3.1.2 GPS(위도,경도)정보

- 원천 데이터에 포함된 GPS의 위치정보는 보수가 필요한 포트홀 및 크랙의 위치를 정확하게 전송할 수 있음
- 주행 안전성 모니터링 시스템을 개발하기 위한 필요한 Database 구축할 수 있는 정보로 활용함
- 원천 데이터에 포함된 GPS정보는 데이터 취득 지역을 파악 할 수 있는 정보이므로 데이터 취득분포 및 수량이 균일화 품질관리 지표로 사용해서 품질관리를 함

2.3.2 획득·정제 절차



2.3.3 획득·정제 기준

2.3.3.1 획득된 원천 데이터 정제

- 취득한 원천 데이터를 육안으로 검수하여 사용할 수 없는 데이터 제거
- 햇빛 반사 등에 의한 밝은 영상, 빗물에 의한 식별 불가 영상
- 신호 대기 등으로 인한 반복적으로 나타나는 동일한 데이터 제거 (GPS, 자동화 도구 등 활용)

<빗물 이미지 및 햇빛 반사 이미지>

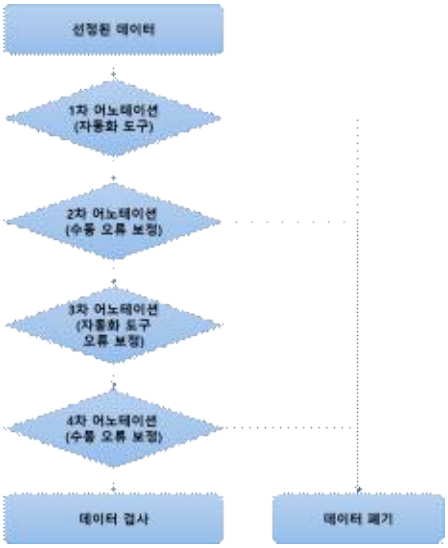
데이터 정제	- 도구를 통하여 자동적으로 영상의 중복성, 동일 위치에 해당하는 데이터 제거 - GPS 주행 속도를 확인하고 이동거리에 따라 후보 데이터 추천 - 수작업을 통하여 중복되거나 변화가 거의 없는 데이터 제거 - 수작업을 통하여 품질이 낮거나 오류가 있는 데이터 정제 - GPS 데이터를 활용하여 지역적 분포를 주기적으로 분석하여 지역적 편향성을 가지지 않도록 정제

2.3.3.2 도구를 활용한 통계 정보 활용을 통한 데이터 편향성 방지

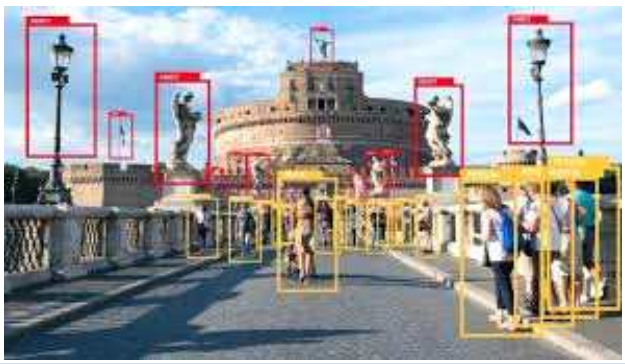
- 모니터링, 리밸런싱 및 추가 데이터 확보 요청, 상황별 분포(지역, 영상기록 시간, 날씨, Class 등), Object 추출 다양성 및 희귀 object 커버리지 모니터링

2.4 어노테이션/라벨링

2.4.1 어노테이션/라벨링 절차



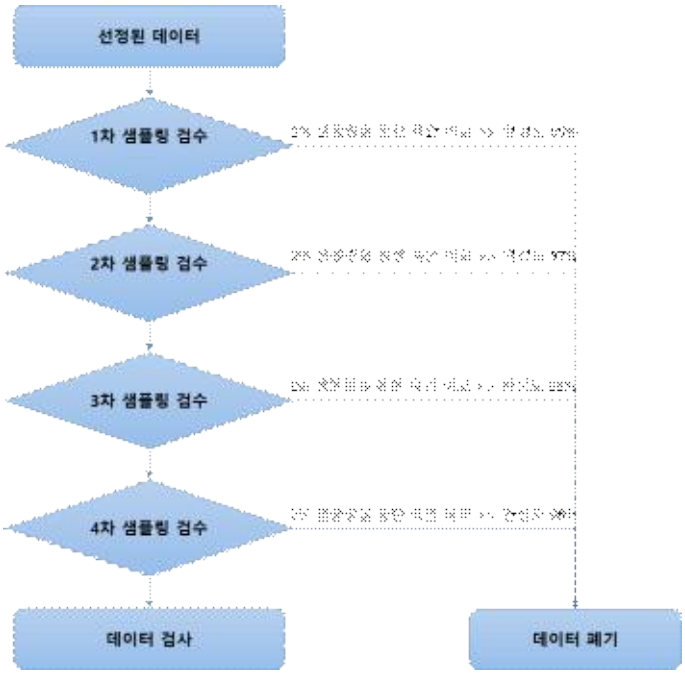
2.4.2 어노테이션/라벨링 기준

	레이블	세부내용
객체 바운딩 박스	동적 객체	▶ 도로상에 출현하는 야생 동물(고라니, 사슴), 보행자 
	정적 객체	▶ 화물차에서의 낙하물(상자, 철근, 판자), 산사태 등의 암석

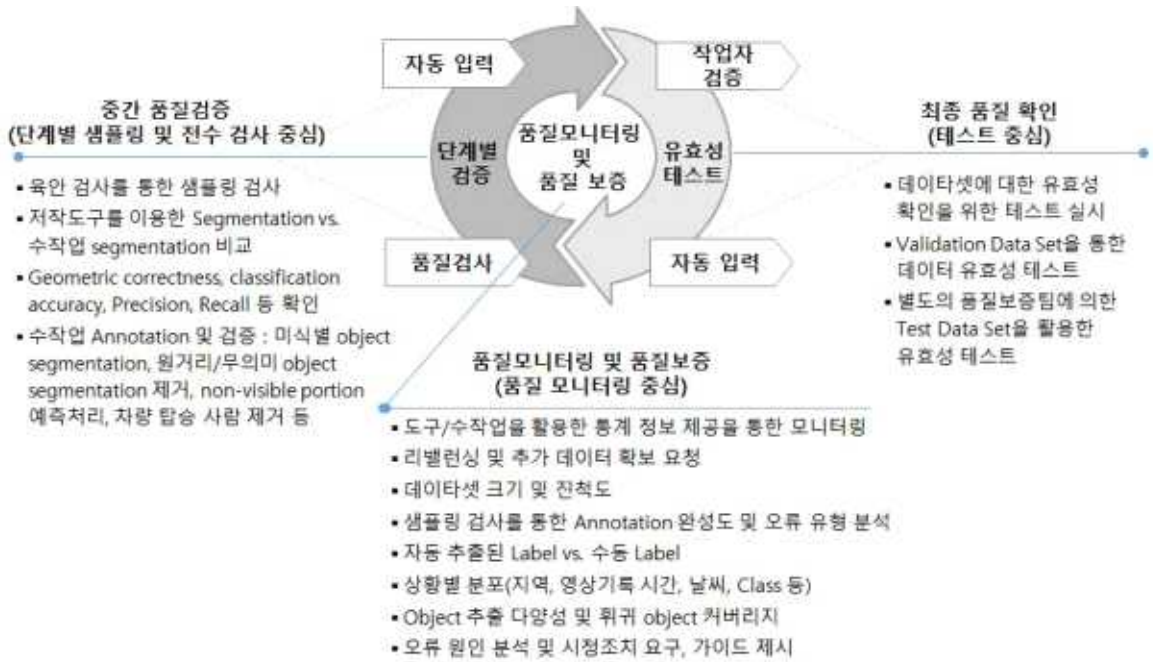
		
	노면	▶ 노면 포트홀 
객체 시멘틱 세그멘테이션	노면	▶ Alligator Crack, Block Crack, Slippage Crack 

2.5 검수

2.5.1 검수 절차



2.5.2 검수 기준



2.6 활용

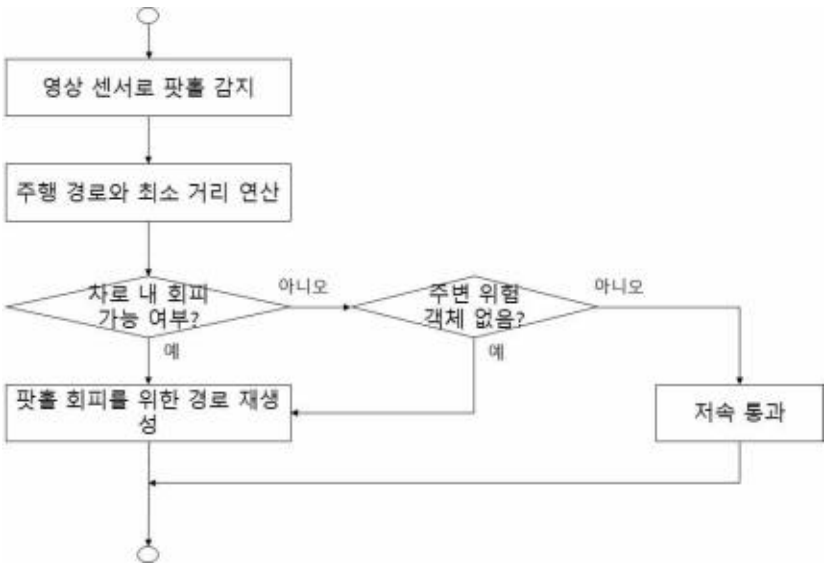
2.6.1 활용 모델

2.6.1.1 모델 학습

- 도로 상의 장애물 및 도로 표면의 이상 상태 검출 및 분할화 네트워크
  - 도로상의 장애물 및 도로 표면의 이상 상태를 검출하기 위하여 Yolo와 같은 검출기 개발을 목표로 함
  - 장면 분할 모델 중 하나인 UPSNet처럼, 장면내에 존재하는 모든 객체들을 분할함을 목표로 함

2.6.1.2 서비스 활용 시나리오

- 도로상태를 활용한 ADAS 개발 및 도로 유지보수를 위한 Logger 시스템 개발
- 자동검출 기반의 위험지역 조기식별을 통한 도로 유지보수 서비스 개발
- 포트홀 감지 후 자율주행 회피 거동 서비스 개발
  - 포트홀 감지 정보를 바탕으로 포트홀 회피를 위한 경로 재생성 알고리즘을 아래와 같은 순서도 기반으로 개발함



2.6.2 데이터 제공

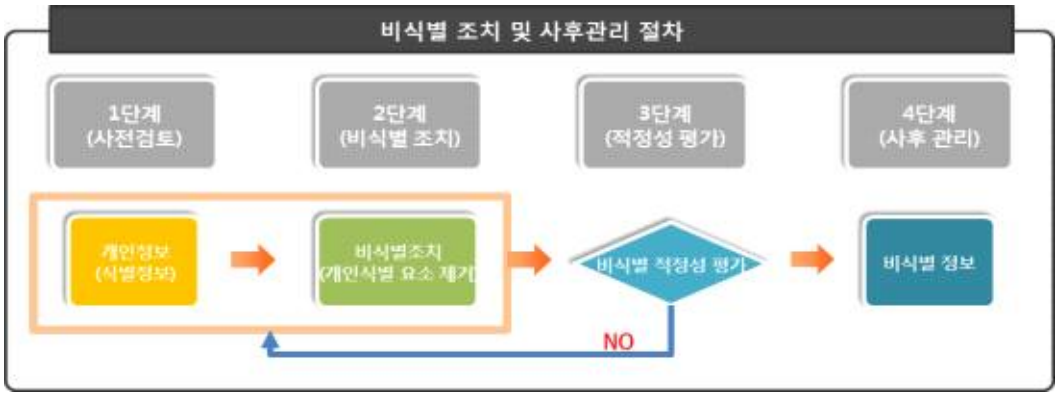
2.6.2.1 구축된 학습용 데이터 자료의 공개·개방 시 저작권은 CC 라이선스를 따름

라이선스	이용조건	문자표기
	• 저작자표시 저작자의 이름, 저작물의 제목, 출처 등 저작자에 관한 표시를 해주어야 함	CC BY
	• 저작자표시-비영리 저작자를 밝히면 자유로운 이용이 가능하지만 영리목적으로 이용할 수 없음	CC BY-NC
	• 저작자표시-변경금지 저작자를 밝히면 자유로운 이용이 가능하지만, 변경 없이 그대로 이용해야 함	CC BY-ND
	• 저작자표시-동일조건변경허락 저작자를 밝히면 자유로운 이용이 가능하고 저작물의 변경도 가능하지만, 2차적 저작물에는 원 저작물에 적용된 것과 동일한 라이선스를 적용해야 함	CC BY-SA
	• 저작자표시-비영리-동일조건변경허락 저작자를 밝히면 이용이 가능하며 저작물의 변경도 가능하지만, 영리목적으로 이용할 수 없고 2차적 저작물에는 원 저작물과 동일한 라이선스를 적용해야 함	CC BY-NC-SA
	• 저작자표시-비영리-변경금지 저작자를 밝히면 자유로운 이용이 가능하지만, 영리목적으로 이용할 수 없고 변경 없이 이용해야 함	CC BY-NC-ND

- 공개된 데이터 활용을 필요로 하는 새로운 산업과 기술 발전으로 개인정보 침해 위험도 증

가 하고 있고, 개인정보 침해가능성을 최소화하면서 데이터의 활용을 높이기 위해 개인정보 비식별화가 필수임

- (개인정보) 살아 있는 개인에 관한 정보로서 개인을 알아볼 수 있는 정보이며, 해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 정보를 포함



<비식별 조치 및 사후관리 순서도>

단계	내용	비고
1단계 사전 검토	개인 정보에 해당하는지 여부 검토후, 개인정보가 아닌 것이 명백한 경우 법제 규제 없이 활용	
2단계 비식별 조치	정보집합물(데이터셋)에서 개인을 식별할 수 있는 요소를 전부 또는 일부 삭제나 대체 등 활용, 개인을 알아 볼 수 없도록 하는 조치	
3단계 적정성 평가	다른 정보와 쉽게 결합하여 개인을 식별할 수 있는지 비식별 조치 적정성 평가	재조취
4단계 사후 관리	비식별 정보 안전조치, 재식별 가능성 모니터링 비식별 정보 활용 과정에서 재식별 방지를 위한 조치	

2.6.2.2 본 사업에서 구축 되는 도로방해물/표면 인지 데이터는 4차 산업혁명의 근간이 되는 AI 학습용 데이터중 하나로 국내 헬스케어, 자율주행 등 AI분야 산업체, 대학, 연구 대상으로 개방

- 국내 대학, 산업체, 연구소가 도로방해물/표면 인지영상 AI 기술 개발을 할 수 있도록 시기적절하게 DB를 공급하여 중소・중견 기업의 머신러닝 기반의 보행 보조 기술개발경쟁력 강화
- 알고리즘 학습이 가능하도록'원본 영상'(추출한 프레임 이미지) 및 '텍스트 파일'(XML 형태 등)이 1개 세트로 구성된 데이터셋 300시간(31,320,000 프레임)을 개방

2.6.2.3 공개·개방의 정책적 기술적 방안

- 정부는 데이터·AI경제 선도 국가로 도약을 위해 데이터·AI경제 활성화 계획 발표
  - 인공지능 혁신생태계를 조성하기 위해 AI 허브*를 구축하고 지속적인 지원 진행 중
    - AI 서비스 개발 필수 인프라인 AI 학습용 데이터, 알고리즘 개발 지원, 컴퓨팅 파워를 온라인 일괄 제공 (http://www.aihub.or.kr)
  - 본 사업에서 생산된 도로상태 및 자율버스 AI 데이터는 AI 학습용 데이터셋 구축 분야 로 AI 허브를 통해 공개·개방하고 관련 소스코드 및 결과 자료는 관련 포럼이나 웹사이트를 통하여 공개하여 관련 사업의 연구 개발자들이 활용할수 있게 지원함

2.6.2.4 도로상태 AI 학습용 데이터 공개를 위한 개방 포털

- 인공지능 학습 데이터는 600시간(수도권 300시간, 광역시 300시간) 이상 구축되는 대용량

데이터로 원활한 공개·개방을 위한 별도의 개방 포털 구성이 필요

- 구축 영상데이터 현황, 자료별 목록, 자료 신청, 다운로드 이력 관리, 게시판 등