Transactions of KSAE, Vol. 32, No. 3, pp.289-294 (March, 2024)

Check for updates

Copyright © 2024 KSAE / 220-05
pISSN 1225-6382 / eISSN 2234-0149
DOI http://dx.doi.org/10.7467/KSAE.2024.32.3.289

자율주행 ODD 확장을 위한 데이터 중심의 AI 개발

정 승 권[*1)] · 전 진 경[1)] · 박 지 은[1)] · 김 재 환[2)]

에이모 DaaS 부문 자율주행팀[1)] · 에이모 DaaS 부문[2)]

A Data-Centric AI Approach to Extend ODD of Autonomous Driving

Seungkwon Jung[*1)] · Jinkyung Jeon[1)] · Jieun Park[1)] · Jaehwan Kim[2)]

[1)]*Autonomous Driving Team, Division of DaaS, AIMMO, 7F, 41 Beolmal-ro, 50beon-gil, Bundang-gu, Seongnam-si, Gyeonggi 13503, Korea*

[2)]*Division of DaaS, AIMMO, 7F, 41 Beolmal-ro, 50beon-gil, Bundang-gu, Seongnam-si, Gyeonggi 13503, Korea*

(Received 13 November 2023 / Revised 20 December 2023 / Accepted 26 December 2023)

Abstract : In January 2023, Mercedes-Benz became the world's first automaker to receive the Level 3 certification as defined by FSAE, and began selling its automated vehicles in the United States in the second half of the year. This development has spurred major automotive manufacturers to accelerate the development of systems that approach Level 3 autonomous driving. As the level of autonomous driving increases, the risk of fatal safety accidents due to environmental perception errors also escalates. To address this issue, securing data that can narrow the gap between real-world and development environments becomes crucial. This paper introduces an environmental perception system for autonomous driving based on the data-centric AI development methodology led by Andrew Ng. Specifically, it proposes a dataset design methodology to expand the Operational Design Domain(ODD), and explains how to implement this effectively through the configuration of a Machine Learning Pipeline after validating its effectiveness.

Key words : Operational design domain(운행 설계 영역), Data-centric AI(데이터 중심의 AI 개발), Machine learning(기계학습), Dataset design(데이터셋 설계), Conditional metrics(조건부 지표)

Subscripts

ODD : operational design domain
SE : scene exposure
IR : input resource
ML : machine learning

1. 서 론

2023년 1월, 메르세데스-벤츠는 FSAE에서 정의한 Level 3 인증을 세계 처음으로 받았고, 그 해 하반기 미국에서 판매를 시작했다. Level 3 차량이 판매되기 시작하면서 완성차 회사들의 자율주행 개발 기술력 경쟁은 가속화되고 있다. 22년 KDB 산업은행에서 발간된 "자율주행차 글로벌 산업 동향"[1)]에서는 2021년 글로벌 판매 비중 1 % 미만이었던 Level 3 자율주행 차량이 2030년에는 50 % 이상이 될 것으로 전망하고 있다.[2)] 향후 Level 3 이상의 자율주행 인증을 받아 도로 주행이 가능한 차량이 많아질 것이다. 자율주행 기술은 빠르게 발전하고 있으나 Level 3에 근접한 자율주행 차량들의 사고는 여전히 발생하고 있으며 테슬라, 웨이모 사례에서 알 수 있듯이 인지 시스템 오류가 높은 비중을 차지한다.

자율주행 레벨이 고도화될수록 개발에 투입해야 하는 시간과 비용은 비약적으로 상승한다. 특히 Level 3 이상의 자율주행 인지 시스템 개발을 위해 확보해야 하는 인공지능 데이터의 양은 방대하다. 정의한 Operational Design Domain(ODD)에 맞춰 성능 목표를 정하고 성능이 취약한 ODD 조건의 성능개선에 집중하면 자율주행 인지 시스템 성능 개선을 위해 투입해야 할 시간과 비용을 줄일 수 있다. 인공지능 시스템 개발에 사용되는 학습데

*Corresponding author, E-mail: ken@aimmo.co.kr
*This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium provided the original work is properly cited.

이터의 설계 단계에서 인공지능 모델의 적용 범위에 맞춰 데이터 유형을 구분하면 선택과 집중을 통해 효율적인 성능 개선이 가능하고 ODD 확장에 투입되는 리소스를 줄일 수 있다.

본 논문에서는 데이터 중심의 AI 개발 방법론을 적용해 자율주행 ODD 확장을 위한 이미지 데이터셋 설계방법을 제안하고 생성된 데이터셋을 활용해 ML Pipeline을 구현해 객체 탐지모델의 ODD 확장에 유효한지 검증한다.

2. 데이터 중심의 AI 개발을 위한 데이터셋 설계

데이터 중심의 AI 개발[3-6]이란 AI 시스템은 코드와 데이터 2개 영역으로 나눌 수 있으며, 코드보다 데이터 개선에 집중하는 것이 AI 시스템의 성능을 향상시키는데 더 효과적이라는 주장이다.

자율주행은 많은 양의 센서 데이터를 다루기 때문에 인지 시스템을 개발할 때 인공지능 모델의 필요에 맞춰 최적화된 데이터를 확보하고 양질의 학습데이터를 공급하는 것이 중요하다.

데이터수집을 진행할 때 현실적인 제약으로 ODD 불균형이 발생할 수밖에 없으며 ODD 조건에 따라 필요한 데이터를 확보했는지 판단할 수 있어야 한다.

ODD 조건 조합으로 데이터의 유형을 구분함으로써 ODD 조건을 고려한 평가지표를 확보할 수 있고 평가지표에 따라 데이터의 양과 질의 개선이 가능해 데이터 중심의 AI 개발이 가능해진다.

2.1 데이터 유형 구분을 위한 인자 정의

ISO 26262[9]에 정의된 “Exposure”의 개념을 해석해 이미지 데이터 기준 “Scene Exposure”라는 지표 계산 방법은 식 (1)과 같다.

$$P(SE) = \frac{N_{condition}}{N_{total}} \tag{1}$$

$P(SE)$는 이미지 데이터 기준 특정 조건의 노출도를 나타내는 확률, $N_{condition}$은 해당 조건에 맞는 이미지 프레임의 개수, N_{total}은 취득된 데이터 프레임의 총 수이다.

데이터를 개선하기 위해 투입해야 하는 ‘Input Resource’는 투입 시간과 비용의 조합으로 표현할 수 있고 지표를 계산하는 방법은 식 (2)와 같다.

$$IR_{time} = \frac{T_{input}}{T_{max}},\ IR_{cost} = \frac{C_{input}}{C_{max}} \tag{2}$$

$$IR_{total} = (0.5 \times IR_{time}) \times (0.5 \times IR_{cost}) \tag{3}$$

IR_{time}은 투입된 시간의 상대적 비중, IR_{cost}는 투입된 비용의 상대적 비중, IR_{total}데이터 개선에 투입된 총 투입 리소스이다. 투입 시간, 투입 비용, 투입 비중은 구축데이터의 속성에 따라 총 투입 리소스의 값이 1이 되도록 설계 담당자가 정의할 수 있다. 할당된 예산, 프로젝트의 목표일정 등으로 정할 수 있다.

2.2 데이터 유형 정의

정의된 인자의 수준을 ISO 26262를 참조해 4단계로 구분하고 각 인자의 조합으로 데이터 유형을 정의해 Table 1, Table 2와 같이 나타냈다.

Table 1 Class of scene exposure

	Class of scene exposure			
	SE1	SE2	SE3	SE4
Description	High	Medium	Low	Very low
P(SE)	> 30 %	over 10~30 %	5 to 10 %	< 5 %

Table 2 Class of input resources

	Class of input resources			
	IR1	IR2	IR3	IR4
Description	Minimal	Moderate	Substantial	Excessive
IR_{total}	< 0.6	0.6 to 0.8	Over 0.8~0.95	> 0.95

Table 3 Data category matrix

Scene exposure class	Input resources class			
	IR1	IR2	IR3	IR4
SE1	Typical	Typical	Typical	Critical
SE2	Typical	Typical	Critical	Critical
SE3	Typical	Critical	Critical	Edge
SE4	Critical	Critical	Edge	Edge

Table 1과 Table 2를 통해 정의한 지표들의 조합으로 Table 3과 같이 데이터 유형이 정리된다. 개발된 모델을 활용해 각 데이터 유형에 대한 평가를 진행하고 부족한 데이터를 수집해 나가면 분할된 조건에 집중해 데이터의 집중 개선이 필요한 영역을 파악할 수 있다.

2.3 ODD 속성과 데이터 유형의 연결

자율주행 차량은 ODD를 기반으로 주행 가능한 영역[7,8]이 정의되기 때문에 Table 3에 정리된 데이터 유형

행렬은 ODD 속성과 연결되어야 한다. 이미지에 많은 영향을 주는 대표적인 날씨 정보와 광원정보를 4단계로 표현하면 아래 Table 4와 같다.

ODD 속성의 각 레벨에 맞춰 SE, IR 지표에 대응시키고 그 조합으로 데이터 유형을 결정할 수 있다. 일관성 있는 데이터 가공 작업을 위해 각 레벨은 수준을 표현할 수 있는 구체적인 문장으로 정의하고 최대한 많은 예시 이미지를 확보해 데이터 가공 작업자들이 직관적으로 인지할 수 있게 해야 한다. 예를 들어 Rainfall extreme은 “폭우로 인해 이미지의 선명도가 낮아지거나, 객체 및 차선의 구분이 어려운 경우”로 표현할 수 있고 Sun extreme은 “태양으로 인해 이미지 전체 밝기에 불균형이 생긴 경우”로 표현할 수 있다. 해당 이미지는 Fig. 1, Fig. 2와 같다.

4단계로 구분된 각 속성의 레벨을 SE1~SE4, IR1~IR4 각 Level의 지표에 대응시킬 수 있고 대응된 값을 기반으로 데이터 유형이 Typical case, Critical case, Edge case중 1개로 선택된다. 모든 ODD 속성값은 각 1개의 데이터 유형에 할당된다.

Table 4 ODD taxonomy attribute expression

ODD		Attribute level			
		Level1	Level2	Level3	Level4
Weather	Rainfall	No	Light	Medium	Extreme
Backlight	Sun	No	Light	Medium	Extreme

Fig. 1 Rainfall extreme image

Fig. 2 Sun extreme image

2.4 ODD 조건부 지표 추출

Table 4에서 ODD 구성요소 각 속성의 수준 정의가 완료됐다면 데이터 유형에 따라 목표 성능지표를 정해 목표 성능지표를 달성하지 못했을 때 데이터를 수집하는 과정을 반복해야 한다. Rainfall을 예시로 데이터 유형을 구분하고 객체 탐지모델을 사용해 목표 성능을 정의하면 Table 5와 같다.

지표를 추출하는 방식은 객체 탐지모델의 대표적인 평가지표인 mAP를 계산하되 평가 대상 이미지의 데이터 유형 추적이 가능하도록 해야 하고 ODD 속성정보를 활용하면 된다. ODD 속성 정보를 사용해 mAP 계산에 필요한 정보를 추출하면 아래 Table 6과 같다. 각 이미지에 존재하는 객체의 추론 결과값은 ODD 속성정보와 데이터 유형정보를 갖는다.

mAP를 계산하는 방법에 데이터 유형 및 ODD구성요소에 따라 필터링하는 단계만 추가해 평가지표를 계산하면 데이터 유형별 평가 결과를 추출할 수 있게 돼 데이터 유형별 목표성능 달성 여부를 확인할 수 있다. ODD 조건에 따라 지표를 추출하는 방법을 전통적인 통계학 용어를 인용해 “ODD 조건부 지표(ODD conditional metrics)”라 정의한다.

Table 5 Conditional metrics based on data category

		Rainfall			
		No	Light	Medium	Extreme
Data category		Typical		Critical	Edge
mAP(Yolov5)	Target	0.9	0.8	0.7	0.6

Table 6 Conditional metrics information

ODD		Label	Inference		
		Class	Confidence	IoU	Class
Rainfall	Extreme	Car	0.956	0.983	Car
		Car	0.938	0.894	Car
		Car	0.906	0.938	Car
		Car	0.571	0.900	Car
		Car	0.187	0.000	Traffic light
	No	Car	0.967	0.96	Car
		Car	0.950	0.905	Car
		Car	0.004	0.015	Pedestrian
		Car	0.003	0.016	Pedestrian

2.5 유효성 검증을 위한 ML Pipeline 설계 및 검증

ODD 조건부 지표 추출이 완료되면 데이터 유형의 각 행렬 요소는 각 ODD 속성 및 데이터 유형을 대표하는 평가지표로 치환될 수 있다. 치환된 평가지표와 비교해 모델의 추론성능이 목표에 미달성 됐다면 달성할 때까지 해당 행렬 요소의 데이터를 수집해 데이터의 양과 질을 개선해 나가는 과정을 반복해야 한다. 이런 과정을 반복하면서 Critical case, Edge case 영역을 줄여 나가면 최소한의 리소스를 투입해 ODD를 확대해 나갈 수 있다. 직관적으로 인지할 수 있도록 그림으로 표현하면 Fig. 3과 같다.

해당 작업을 반복하는 과정은 수동으로 진행할 수 없으므로 Airflow, Kubeflow, ClearML 같은 ML 자동화 툴을 사용해 반복작업 파이프라인을 구성해야 한다.

ClearML을 사용해 간단한 파이프라인을 구성해 설계된 데이터셋을 사용해 자동화 파이프라인 구현이 가능한지 검증했다. ODD 조건부 지표와 비교해 추론결과가 목표 성능에 마달된 경우 오토트리거 설정을 통해 데이터를 자동으로 가져와 학습, 검증, 평가를 반복하는 자동화 파이프라인을 구현할 수 있었다. 검증한 ClearML 파이프라인 컨셉은 Fig. 4와 같다.

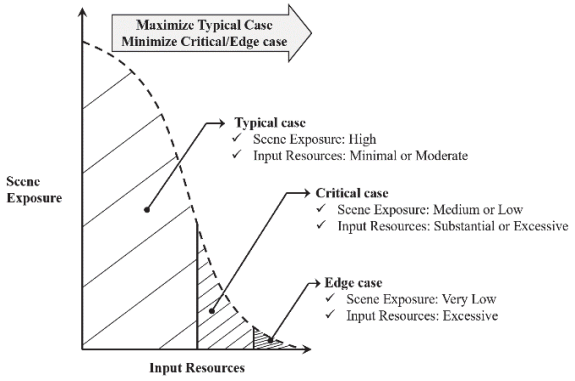

Fig. 3 ODD Extension visualization

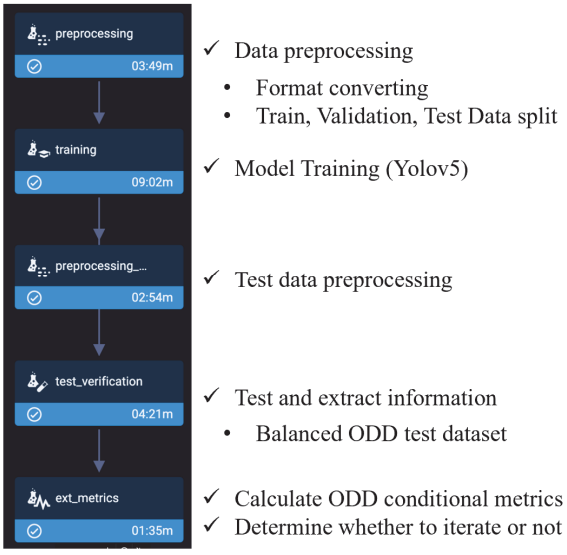

Fig. 4 ClearML pipeline

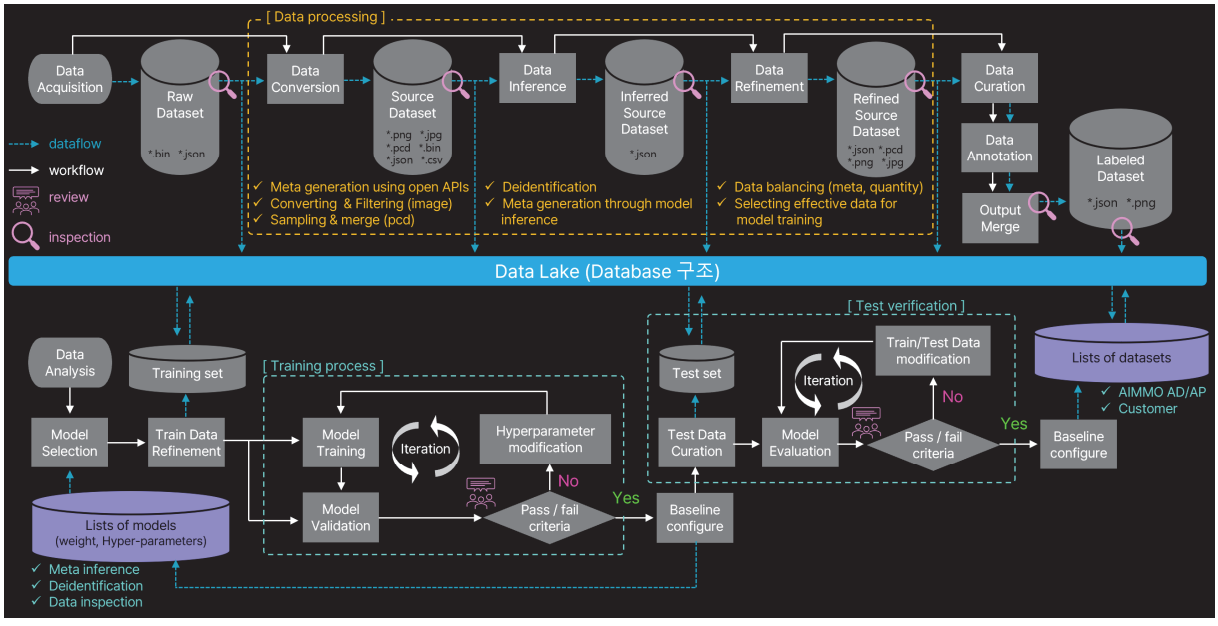

Fig. 5 Data-Centric AI ML pipeline concept design

데이터 중심의 인공지능 시스템을 개발하려면 데이터 생산 전주기 영역을 통제할 수 있는 파이프라인 컨셉을 설계해야 한다. 검증된 유효성에 기반해 데이터 수집, 변환, 처리, 가공, 검수, 학습, 검증, 평가 등 ML 전주기 개발 Cycle을 고려해 구체화하면 Fig. 5와 같다.

설계된 파이프라인의 컨셉은 자동화뿐만 아니라 데이터 베이스 구조의 데이터 레이크와 연동돼 수집된 데이터의 반복적인 개선을 자동화할 수 있는 구조이다.

인공지능 시스템 개발은 평가를 반복[10]하면서 모델의 안정화, 고도화를 진행한다. 개발 Cycle이 반복될 때 모델을 고정하고 데이터를 개선하거나, 데이터를 고정하고 모델을 개선하는 방식 중 효과적인 방법을 선택해 변경인자의 영향을 최소화한 상태로 1 Cycle 개발을 완료해야 한다. 결과에 따라 다음 개발 Cycle을 진행할 때 두 개발 방식 중 효과적인 방식을 취할 수 있어 한 가지 개발방식만 택하는 것 대비 모델 성능의 한계지점을 높일 수 있도록 ML파이프라인을 설계했다.

제안한 방법으로 객체 탐지모델을 사용해 대량의 데이터를 학습시키고 최적화시켜 나가는 과정을 수행한 결과 Fig. 6과 같은 실험 결과를 얻었다.

가로축은 객체의 클래스별 학습데이터 이미지 개수, 좌측 세로축은 객체 개수, 우측 세로축은 객체 탐지모델의 성능 지표 및 정규화된 투입 리소스를 의미한다. 범례에는 총 투입 리소스와 ODD 구성요소(시간대, 날씨)의 평가지표를 볼 수 있도록 표현했고 Resource는 IR 지표와 동일한 의미이다.

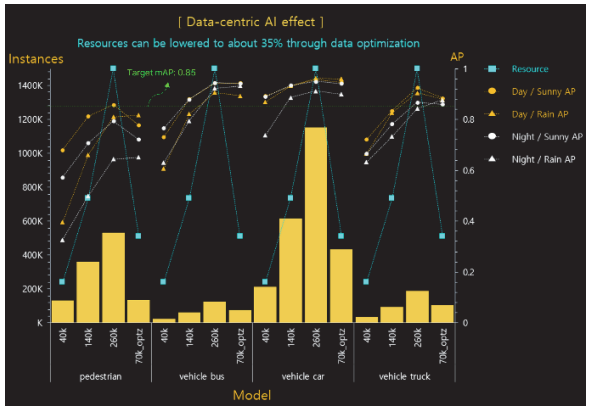

Fig. 6 Data-Centric AI effect

Table 7 Input resource

Model	Input resources	
	Training time(hours)	Data storage(GB)
40 K	240	101
140 K	420	332
260 K	1,688	558
70 K_optz	400	232

4만(40 K), 14만(140 K), 26만(260 K) 프레임으로 양을 늘려 가면서 학습한 결과 학습량이 늘어남에 따라 성능이 높아지는 것을 확인할 수 있고 26만 프레임과 최적화된 7만 프레임(70 k_optz) 학습결과 모두 최초 성능목표인 “Target mAP: 0.85”를 달성했다.

정의한 ODD 조건인 시간대(주간/야간)조건 및 날씨(맑음/비) 조건의 데이터 유형을 구분해 지표를 추출하고 성능이 떨어지는 데이터 유형의 양과 질을 개선해 가면서 학습한 결과 데이터 생산에 투입되는 리소스를 35 % 수준까지 줄일 수 있었다.

비용을 제외한 데이터 생산에 투입되는 리소스 계산은 Table 7과 같다. 최대 리소스가 투입된 260 K의 지표를 1로 정규화해 계산한다.

실험데이터는 에이모에서 자체 구축한 AD Dataset ver 1.0을 사용했다. 해당 결과는 데이터를 구축해 가며 후행적으로 얻은 결과지만 한번 수행해 얻은 ODD 조건부 특성 지표는 신규 구축하고자 하는 자율주행 데이터 유형에 적용해 일반화 가능하다.

3. 결 론

1) 데이터 구축 초기 단계에서 ODD 속성과 데이터의 유형 구분을 통해 데이터를 설계하는 방법을 제안함으로써 데이터 가공작업의 주체인 인간이 자율주행 도메인에 특화된 데이터의 특징을 인지할 수 있는 직관성이 높아졌다. 직관성이 높아짐에 따라 데이터 일관성 확보에 유리해져 인공지능 시스템 개발 효율성이 좋아졌다.
2) 제안한 방법으로 데이터셋을 설계함으로써 많은 리소스 투입을 해야 하는 데이터셋 튜닝과 모델 튜닝과정의 자동화 설계가 가능해졌다. 자동화 설계를 통해 인공지능 시스템의 개발 및 배포기간이 단축될 수 있고 투입 리소스 절감으로 이어진다.
3) 모델이 꼭 필요로 하는 분할된 데이터 영역에 맞춰 데이터 개선 및 성능평가가 가능해져 불필요한 데이터 수집이 줄어들고 투입 리소스 절감으로 이어진다.
   본 연구에서 정의한 지표 SE, IR과 데이터 유형을 결정하는 방식은 도메인 전문지식을 바탕으로 구축하고자 하는 인공지능 데이터에 대한 이해가 수반되어야 한다.
   정의해야 하는 정량지표는 전문가의 경험과 데이터에 근거해 수립되어야 한다. 경력있는 연구자와 함께 리뷰하는 과정을 필수로 진행한다면 인공지능 시스템 개발의 효율화 성과는 더 커질 것이다.

후 기

본 연구는 국토교통부/국토교통 과학기술 진흥원의 지원으로 수행된 연구임(과제번호 RS-2022-00142565).

References

1) S. Park, Global Industrial Trends for Self-Driving Vehicles, KDB Korea Development Bank's Monthly Report, Vol.801, pp.50-66, 2022.
2) J. Kim and S. -C. Kee, "A Research on the ODD and OEDR Guidelines Based on the Demonstration Case of Autonomous Driving in Sejong City," Transactions of KSAE, Vol.28, No.10, pp.659-668, 2020.
3) Q. Lin, G. Ye, J. Wang and H. Liu, "RoboFlow: A Data-Centric Workflow Management System for Developing AI-enhanced Robots," 5th Conference on Robot Learning, pp.1789-1794, 2022.
4) M. H. Jarrahi, A. Memariani and S. Guha, "The Principles of Data-Centric AI," Communications of the ACM, Vol.66, No.8, pp.84-92, 2023.
5) D. Alvarez-Coello, D. Wilms, A. Bekan and J. Gómez, "Towards a Data-Centric Architecture in the Automotive Industry," Procedia Computer Science, Vol.181, pp.658-663, 2021.
6) Y. Lee, O. J. Kwon, H. Lee, J. Kim, Kangwook Lee and K. -E. Kim, "Augment & Valuate: A Data Enhancement Pipeline for Data-Centric AI," 35th Conference on Neural Information Processing Systems, 2021.
7) ISO_DIS_34503(en) - Road Vehicles-Test scenarios for automated driving systems-Taxonomy for operational design domain.
8) ISO_21448_2022(en) - Safety of the intended functionality.
9) ISO_26262-3 - Road vehicles — Functional safety — Part 3: Concept phase.
10) E. Thorn, S. C. Kimmel, M. Chaka and B. A. Hamilton, A Framework for Automated Driving System Testable Cases and Scenarios(No. DOT HS 812 623), U.S. Department of Transportation, National Highway Traffic Safety Administration, 2018.