

노이즈 섭동을 통한 환경 변화에 강건한 카메라 기반 3 차원 점유 예측

서정현, 송승현, 이재구*

국민대학교 컴퓨터공학과

*jaekoo@kookmin.ac.kr

요약

3 차원 점유 예측은 자율 주행 인지 과업에서 차량 주변의 객체를 비롯한 전반적인 주변 환경을 이해하는데 있어 필수적인 과업이다. 이때, 안전한 주행을 위해서는 날씨 변화와 카메라의 이상 등과 같은 예상하지 못한 환경 변화에서도 강건한 모델을 만들어야 한다. 따라서, 본 논문은 먼저 기존의 학습된 모델을 변형 데이터 집합으로 평가하는 경우에 원본 이미지와 대비하여 성능 저하가 발생함을 확인했다. 또한, 이를 해결하기 위해 학습 시에 이미지 특징에 노이즈 섭동을 적용함으로써 환경 변화에도 강건한 3 차원 점유 예측 모델을 만드는 방법을 제안한다. 우리는 실험을 통해 기존의 학습 방법에 비해 mIoU 가 최대 약 14.78% 향상된 것을 보이며 해당 방법이 유의미함을 입증했다.

1. 서론

최근 딥러닝(Deep Learning)을 자율 주행의 다양한 과업에 적용하고자 하는 연구가 활발히 이루어지고 있다[1, 2]. 특히, 자율 주행 인지 과업에서는 딥러닝 모델을 활용하여 다중 시점(Multi-view)의 카메라로부터 촬영한 이미지들을 조감도(Bird's-Eye-View, BEV)의 형태로 변환함으로써 차량 주변 환경에 대한 전반적인 이해를 가능하게 한다.

자율 주행 인지 과업 중 하나인 3 차원 점유 예측(3D Occupancy Prediction)은 입력 이미지를 바탕으로 차량 주변 객체들의 점유 여부와 해당하는 클래스 정보를 3 차원의 복셀 그리드(Voxel Grid) 형식으로 예측하는 과업이다. 또 다른 인지 과업 중 하나인 3 차원 객체 탐지(3D Object Detection)는 객체 중심의 위치를 분석하는 반면에, 3 차원 점유 예측은 공간의 전체적이고 세밀한 정보를 파악할 수 있다는 점에서 장점을 보이며 활발히 연구되고 있다[3].

전통적으로 딥러닝 모델은 학습 데이터와 다른 분포의 입력 데이터를 마주하면 분포 변화(Distribution Shift)로 인해 성능 저하가 발생하는 문제점이 존재한다. 이를 해결하고자 도메인 일반화(Domain Generalization)와 같은 연구들이 등장하였고 [4], 정규화 방법[5, 6]을 통해 강건한 모델을 만드는 연구가 이루어지고 있다. 특히, 자율 주행 상황에서는 날씨 변화 및 카메라의 이상 등과 같은 환경 변화가 빈번히 발생하기 때문에 안전한 도로 주행을 위해서는 다양한 환경 변화에도 성능 저하 없이 강건한 모델을 만드는 것이 중요하다.

따라서, 우리는 다양한 환경 변화에도 강건한 3 차원 점유 예측 모델을 만들기 위한 방법을 제시하고자 한다. 먼저, 기존의 데이터 집합을 [그림 1]과 같은 변형이 가해진 데이터 집합[7]으로 만들고, 모

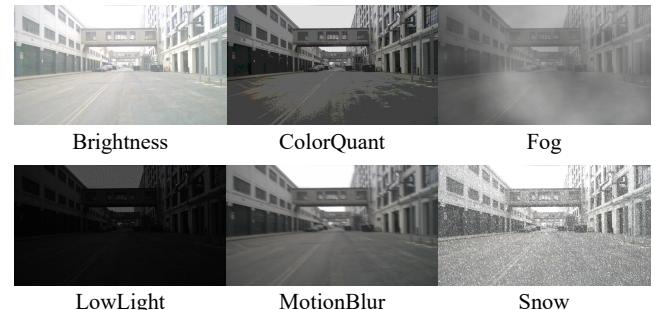


그림 1. 변형 데이터 집합 nuScenes-c[7]의 예시

델에 평가함으로써 환경 변화에 따른 성능 저하가 발생함을 확인하였다. 또한, 이를 해결하고자 우리는 학습 시에 이미지 백본(Image Backbone)으로부터 얻은 특징(Feature)에 정규화를 적용하는 과정에서 노이즈 섭동(Noise Perturbation) 더해줌으로써 다양한 분포에 대해 학습할 수 있도록 했다. 이때 학습 가능한 파라미터(Parameter)가 아닌 랜덤 노이즈를 사용함으로써 추가적인 연산 과정 없이 효율적인 학습이 가능하게 했고, 평가 시에는 과적합을 막기 위해 노이즈를 사용하지 않았다. 이를 통해 우리는 여러 환경 변화 데이터 집합에서 더 나은 일반화 성능을 달성하며 우리의 실험이 효과적임을 보인다.

2. 관련 연구

2.1 카메라 기반 3 차원 점유 예측

카메라 기반 3 차원 점유 예측 과업은 차량을 중심으로 주변에 존재하는 객체의 클래스 정보를 복셀마다 예측하는 과업이다. 해당 과업은 (i) 입력 이미지로부터 특징들을 추출하는 과정, (ii) 2 차원의 이미지 특징들을 3 차원의 복셀 특징으로 변환하는 과

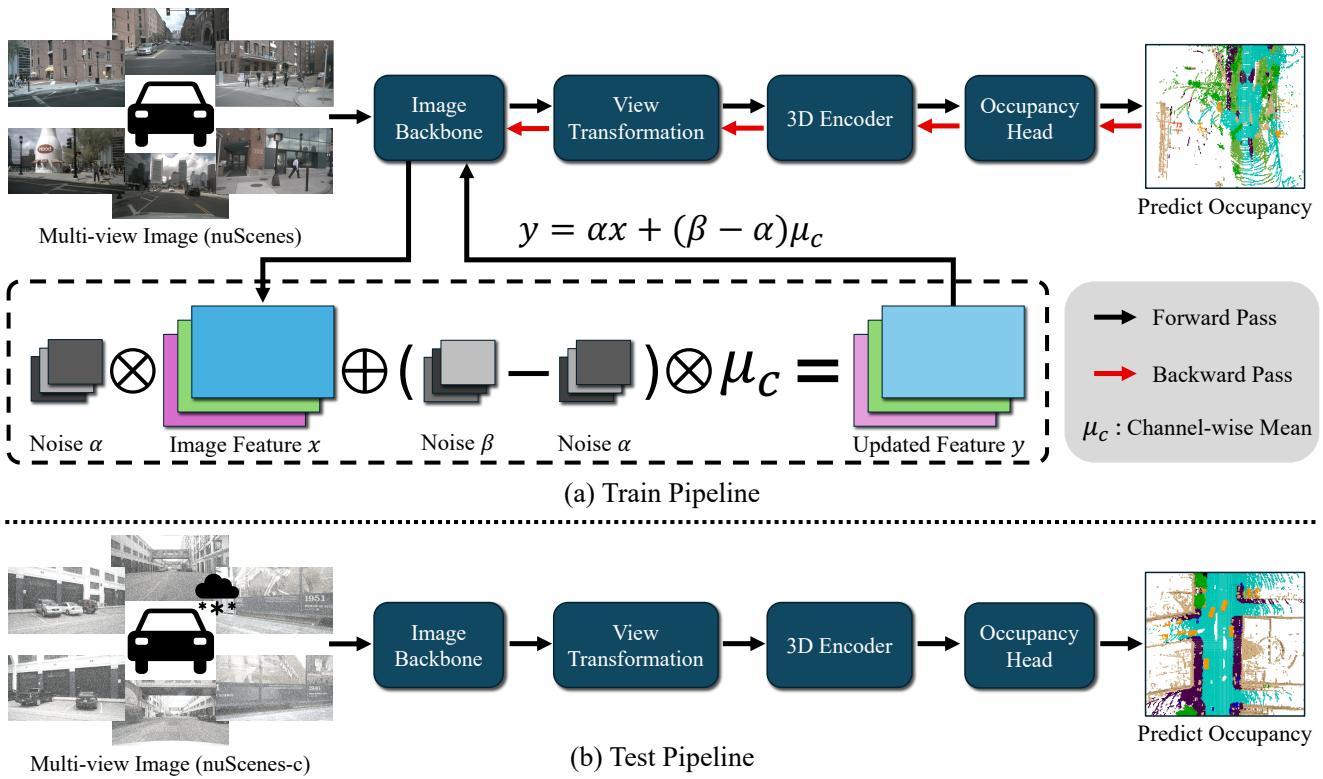


그림 2. 학습 및 평가 구조도

정, (iii) 3 차원의 복셀 특징으로부터 유의미한 정보를 추출하는 과정, (iv) 일정 크기의 조감도 공간 상에서 복셀마다 점유 여부와 클래스를 예측하는 과정으로 구성되어 있다.

FB-OCC[8]는 2 차원의 이미지 특징을 3 차원의 복셀 특징으로 시점 변환하는 과정에서 전방 투영(Forward-projection)과 후방 투영(Backward-projection)을 함께 진행하여 정보 손실 없이 효과적으로 복셀 특징을 추출하도록 한 연구이다.

PanoOcc[9]는 복셀 쿼리(Voxel Query)를 기반으로 이미지 특징과 어텐션(Attention) 연산을 통해 시점을 변환하는 방법과 이전 프레임의 정보를 활용하는 시간적 인코더(Temporal Encoder)를 사용함으로써 복셀 쿼리가 3 차원 공간 내에서 효과적인 특징을 가질 수 있도록 한 연구이다.

우리는 앞서 소개한 3 차원 점유 예측 과업에서의 대표 모델인 FB-OCC[8]와 PanoOcc[9]에 대해 환경 변화에 따른 성능 저하를 확인하고, 강건한 모델을 만드는 방법을 적용하여 기존의 학습 방법과 비교하고자 한다.

2.2 정규화 방법을 통한 강건한 학습

정규화(Normalization) 방법은 딥러닝 모델의 일반화 성능을 높이기 위해 자주 사용되는 방법 중 하나이다. 특히, 특징 수준(Feature Level)에서의 정규화 방법은 통계 값인 평균과 분산을 활용하여 특징 값을 조절함으로써 효과적인 학습을 가능하게 한다.

AdaIN(Adaptive Instance Normalization)[10]은 이미지에서의 스타일 전이(Style Transfer) 과업을 위해 제

안된 방법이다. 해당 연구는 수식 (1)과 같이 정규화 과정에서 스타일 특징과 컨텐츠 특징의 평균과 분산을 활용하여 과업을 수행한다.

$$y = \sigma_s \left(\frac{x - \mu_c}{\sigma_c} \right) + \mu_s \quad (1)$$

μ_s 과 σ_s 는 스타일 특징의 평균과 분산을, μ_c 와 σ_c 는 컨텐츠 특징의 평균과 분산을 의미한다.

MixStyle[5]은 이미지 분류(Image Classification) 과업에서 강건한 모델을 만들고자 한 연구이다. 해당 연구는 입력 이미지로부터 얻은 특징 정보에 섭동을 주는 방식의 정규화를 적용함으로써 모델이 다양한 스타일에 대해 학습하도록 한다.

NP(Normalization Perturbation)[6]은 2 차원 객체 탐지(2D Object Detection) 과업에서 일반화 성능을 높이기 위한 방법을 제안한 연구이다. 해당 연구는 이미지 특징에 랜덤 노이즈를 더하는 방식으로, 입력 데이터와는 무관한 본질적인 정보에 주목함으로써 도메인 불변(Domain Invariant) 특징을 학습하도록 한다.

MixStyle[5]과 NP[6] 모두 AdaIN[10]의 수식(1)을 기반으로 하여 특징 수준에서의 정규화 방법을 통해 일반화 성능을 높여 강건한 모델을 만드는 것을 목표로 한다. 앞선 방법과 유사하게, 본 논문 역시 정규화 과정에서 노이즈 섭동을 통해 강건한 3 차원 점유 예측 모델을 만드는 방법을 제안한다.

3. 본론

3 차원 점유 예측 과업은 현재의 시점 t 에서, N 개의 시점 이미지 $\{I_1, I_2, \dots, I_N\}$ 을 입력으로 받아 차

량을 중심으로 3 차원의 복셀 공간 (X, Y, Z)를 예측하는 과업이다. 이때 n 개의 클래스를 예측해야 하는 경우에 각각의 복셀을 $\{c_0, c_1, \dots, c_n\}$ 과 같이 표현할 수 있고, 이때 c_0 는 해당 복셀에 물체가 존재하지 않아 비어 있는 상태를 의미한다.

[그림 2]는 제안하는 학습 (a) 및 평가 (b) 방법에 대한 전반적인 개요를 나타낸다. 우리는 환경 변화에 강건한 모델을 만들고자 카메라 입력과 직접적인 관계가 있는 이미지 백본에서 도메인 불변(Domain Invariant) 특징을 추출하고자 했다. 앞서 이미지 백본으로부터 얻은 특징은 [그림 2]와 같이 시점 변환(View Transformation), 3 차원 인코더(3D Encoder), 점유 헤드(Occupancy Head)를 거쳐 최종적으로 조감도(BEV) 형태의 예측 결과를 얻는다.

$$y = \sigma_s \left(\frac{x - \mu_c}{\sigma_c} \right) + \mu_s, \quad \sigma_s = \alpha \sigma_c, \quad \mu_s = \beta \mu_c \quad (2)$$

수식 (2)는 모델로부터 얻은 특징에 노이즈 섭동을 적용하는 과정을 나타내고 있다. 이때 α 와 β 는 가우시안(Gaussian) 분포의 노이즈에 해당하고, μ_c 와 σ_c 는 각각 특징의 채널 별 평균과 분산에 해당한다.

$$y = \alpha x + (\beta - \alpha) \mu_c \quad (3)$$

수식 (2)를 전개하면 수식 (3)과 같이 채널 별 평균만을 활용하여 노이즈 섭동 과정을 나타낼 수 있고, [그림 2]의 (a)에서와 같이 학습시에만 노이즈 섭동을 적용하여 특징을 업데이트 한다.

우리는 얕은 레이어(Shallow Layer)에서의 노이즈 섭동이 더 나은 일반화된 모델을 학습한다는 점[6]을 바탕으로 이미지 백본의 첫 번째, 두 번째 레이어로부터 얻은 특징에만 노이즈 섭동을 적용했다. 추가적으로, 과도한 노이즈로 인해 학습이 저하되는 것을 막고자 확률 값이 0.5 이상인 경우에만 노이즈 섭동을 더했다. 또한, 지역적인 정보와 전역적인 정보를 모두 얻고자 FPN(Feature Pyramid Network)[11] 방식을 통해 각 레이어로부터 나온 특징을 종합하여 최종 특징으로 사용한다. 앞선 방법으로 이미지 백본으로부터 최종 특징을 얻고, 나머지 모듈들을 통과하며 3 차원 복셀을 예측하는 방향으로 학습이 진행된다. 노이즈 섭동 방식은 [그림 2]의 (a)와 같이 학습 시에만 진행되고, 평가 시에는 과적합을 방지하고자 [그림 2]의 (b)와 같이 노이즈 섭동 없이 변형된 데이터 집합으로 평가를 진행한다.

4. 실험 및 결과

우리의 방법은 이미지 백본으로부터 얻은 특징에 노이즈 섭동을 적용하는 방법으로 모델과 무관한 플러그-앤-플레이(Plug-and-Play) 방식이다. 따라서, 우리는 3 차원 점유 예측 과업의 대표적인 모델인 FB-OCC[8]와 PanoOcc[9]에 대해 각각 실험을 진행했다. 두 모델 모두 이미지 특징 추출을 위해 주로 사용되는 CNN 계열의 ResNet[12]을 백본으로 사용하며, 실험은 기존 모델[8, 9]에서의 방법을 따랐다.

학습 데이터 집합은 자율주행의 여러 과업에서 대표적으로 사용되는 데이터 집합인 nuScenes[13]를 사용했다. nuScenes[13]는 하나의 프레임에 대해 전방, 후방, 측면으로부터 촬영한 6 개의 시점의 이미지로 이루어져 있어, 다중 카메라 기반 자율 주행 인지 과업에서 주로 사용하는 데이터 집합이다. 평가 데이터 집합은 학습한 모델의 강건함을 평가하고자 nuScenes[13]에 여러 변형을 가한 nuScenes-c[2]를 사용했다. 이때, 변형의 종류에는 Snow, Fog 등이 있고, easy, mid, hard 와 같이 다양한 강도의 변형된 이미지에 대해 평가했다. 평가 지표로는 3 차원 점유 예측 과업에서 주로 사용하는 mIoU(mean Intersection over Union)를 사용했다.

[표 1]은 여러 변형을 가한 이미지에 대해 평가한 결과를 나타내고 있다. 기존 방법 대비 우리의 방법으로 학습한 경우에 더 나은 성능을 보여주는 것을 확인할 수 있고, 특히, 변형의 정도가 어려운 경우에 mIoU 가 최대 약 14.78% 향상했음을 확인할 수 있다. 이는 노이즈 섭동이 정규화 과정에서 도메인과 무관한 다양한 분포를 학습하도록 하여 강건한 모델을 만들 수 있음을 의미한다.

[그림 3, 4, 5]는 점유 예측 결과를 시각화한 것을 나타내고 있다. [그림 3]은 nuScenes[13]에 존재하는 6 개의 카메라 중 전방 카메라로부터 얻은 이미지에서의 결과를 나타낸다. 입력 이미지를 보면, 변형 정도가 어려운 이미지임에도 불구하고 기존 모델 대비 더 나은 점유 예측 결과를 보여준다. [그림 4]는 (200x200x16) 크기를 갖는 조감도(BEV)에서의 결과를, [그림 5]는 실제 자율 주행 시에 사용할 수 있는 노말 시점(Normal View)에서의 점유 예측 결과를 나타낸다. 결과를 보면, 우리의 방법을 통해 학습시킨 모델이 기존 모델에 비해 정답 라벨인 GT 와 더 유사한 결과를 보이는 것을 알 수 있다. 이는, 노이즈를 섭동을 적용하여 학습시키는 것 만으로도 환경 변화에 강건한 모델을 만들 수 있음을 보인다.

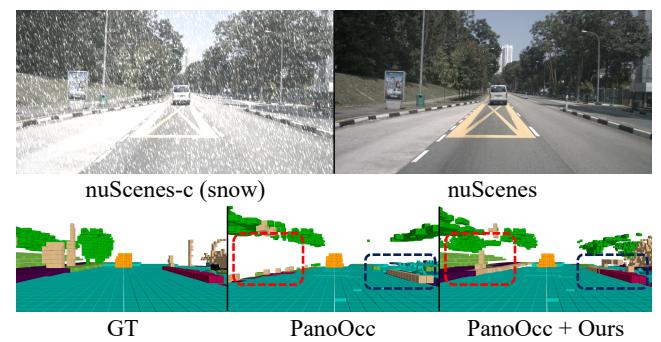


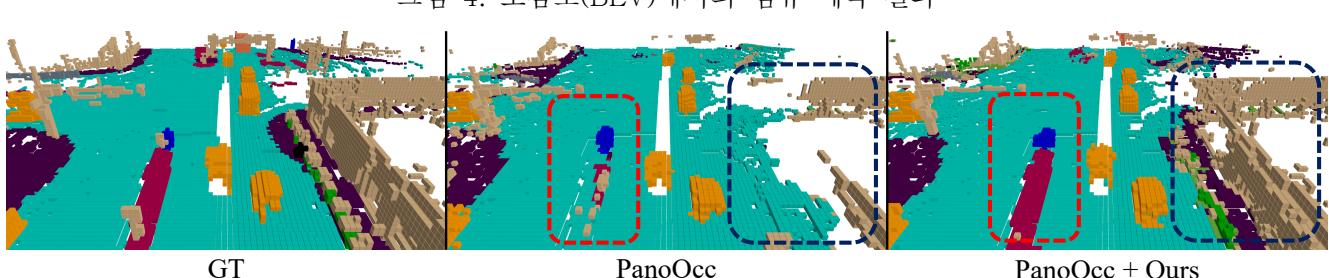
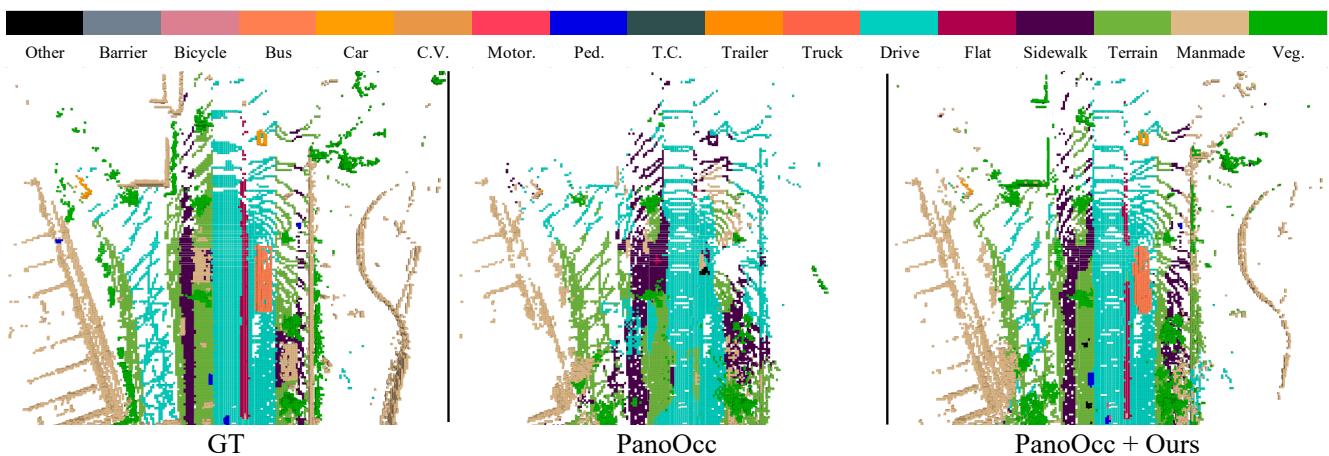
그림 3. 전방 시점에서의 이미지 및 점유 예측 결과

5. 결론

자율주행 상황에서는 날씨 혹은 카메라의 이상과 같은 다양한 환경 변화가 발생하고, 이때 학습된 모델이 예상하지 못한 상황에도 효과적으로 동작하는 것은 원활한 주행을 위해 필수적이다. 따라서,

표 1. nuScenes-c[7] 데이터 집합에서의 평가 결과

Method \ Corruption	mIoU (%; ↑)					
	Brightness	ColorQuant	Fog	LowLight	MotionBlur	Snow
nuScenes-c (easy)						
FB-OCC [8]	34.62	35.98	29.72	21.00	32.63	13.11
FB-OCC + Ours	35.97	35.94	31.98	24.50	35.94	22.08
PanoOCC [9]	35.96	35.82	33.38	30.40	32.92	19.52
PanoOCC + Ours	34.84	34.72	33.09	31.36	34.42	28.62
nuScenes-c (mid)						
FB-OCC [8]	26.89	26.35	27.85	17.53	20.88	5.87
FB-OCC + Ours	32.93	31.98	29.75	19.06	31.22	12.96
PanoOCC [9]	34.36	31.59	32.36	26.18	22.60	12.26
PanoOCC + Ours	34.13	32.32	32.56	28.74	32.41	25.80
nuScenes-c (hard)						
FB-OCC [8]	22.04	13.05	24.78	12.09	17.30	4.95
FB-OCC + Ours	30.36	22.18	26.89	12.38	28.02	12.35
PanoOCC [9]	32.53	23.58	29.93	19.92	18.28	9.34
PanoOCC + Ours	33.46	26.90	31.00	23.81	31.40	24.12



감사의 글

본 연구는 과학기술정보통신부의 재원으로 한국연구재단의 지원(No.RS-2023-00212484; 복잡한 실제 주행환경에서 설명 가능한 움직임 예측)과 정보통신기획평가원의 생성 AI 선도인재양성사업(IITP-2024-RS-2024-00397085) 연구 결과로 수행되었음

참고문헌

- [1] Li, Zhiqi, et al. "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers." *European conference on computer vision*. Cham: Springer Nature Switzerland, 2022.
- [2] Peng, Lang, et al. "Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023.
- [3] https://youtu.be/jPCV4GKX9Dw?si=BxElIc1wDVA_XbyTF, Tesla AI DAY 2022
- [4] Choi, SungHa, et al. "Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
- [5] Zhou, Kaiyang, et al. "Domain generalization with mixstyle." *arXiv preprint arXiv:2104.02008* (2021).
- [6] Fan, Qi, et al. "Normalization perturbation: A simple domain generalization method for real-world domain shifts." *arXiv preprint arXiv:2211.04393* (2022).
- [7] Xie, Shaoyuan, et al. "Benchmarking and Improving Bird's Eye View Perception Robustness in Autonomous Driving." *arXiv preprint arXiv:2405.17426* (2024).
- [8] Li, Zhiqi, et al. "Fb-occ: 3d occupancy prediction based on forward-backward view transformation." *arXiv preprint arXiv:2307.01492* (2023).
- [9] Wang, Yuqi, et al. "Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024.
- [10] Huang, Xun, and Serge Belongie. "Arbitrary style transfer in real-time with adaptive instance normalization." *Proceedings of the IEEE international conference on computer vision*. 2017.
- [11] Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- [12] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- [13] Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.