Fisheye 변환을 활용한 InternImage 기반의 도메인 적응 이미지 분할

2023 Samsung AI Challenge : Camera-Invariant Domain Adaptation

2023.10.06

팀명 : GNOEYHEAT

팀원 : 김태형, 박세홍, 허준호

주최 SAMSUNG 주관 DACON

목차

1. 문제 정의

- 왜곡이 존재하지 않는 이미지(Source Domain)와 레이블을 활용하여, 왜곡된 이미지(Fisheye Target Domain)에 대해서도 고성능의 이미지 분할(Semantic Segmentation)을 수행하는 AI 알고리즘을 개발하고자 함.
- Domain Adaptive Semantic Segmentation 문제로 정의함.
- Datasets
  - train_source_image / train_source_gt
    - 총 2,194장의 2048 x 1024 크기의 학습 데이터 이미지 / 픽셀값 0~11, 255(배경)으로 구성된 Ground Truth 이미지
  - train_target_image
    - 총 2,923장의 1920 x 1080 크기의 학습 데이터 이미지 (Fisheye 형태의 Target 이미지)
  - val_source_image / val_source_gt
    - 총 466장의 2048 x 1024 크기의 검증 데이터 이미지 / 픽셀값 0~11, 255(배경)으로 구성된 Ground Truth 이미지
  - test_image
    - 총 1,898장의 1920 x 1080 크기의 평가 데이터 이미지 (Fisheye 형태의 Target 이미지)
  - sample_submission.csv
    - 960 x 540으로 조정된 이미지를 사용하여 Run Length Encoding(RLE)로 표현된 이진마스크 정보를 가짐.

2. 탐색적 데이터 분석

데이터 예시

- Source 데이터는 왜곡이 없는 (Rectilinear Source Domain) 이미지, Target 데이터는 왜곡된 (Fisheye Target Domain) 이미지임.
- 학습에 사용할 수 있는 레이블이 존재하지 않는 왜곡된 이미지가 존재함.
- Ground Truth의 범주는 총 13개로 아래와 같음.
  - 0~11 ('Road', 'Sidewalk', 'Construction', 'Fence', 'Pole', 'Traffic Light', 'Traffic Sign', 'Nature', 'Sky', 'Person', 'Rider', 'Car'), 255 ('Background')

TRAIN_SOURCE_0000.png

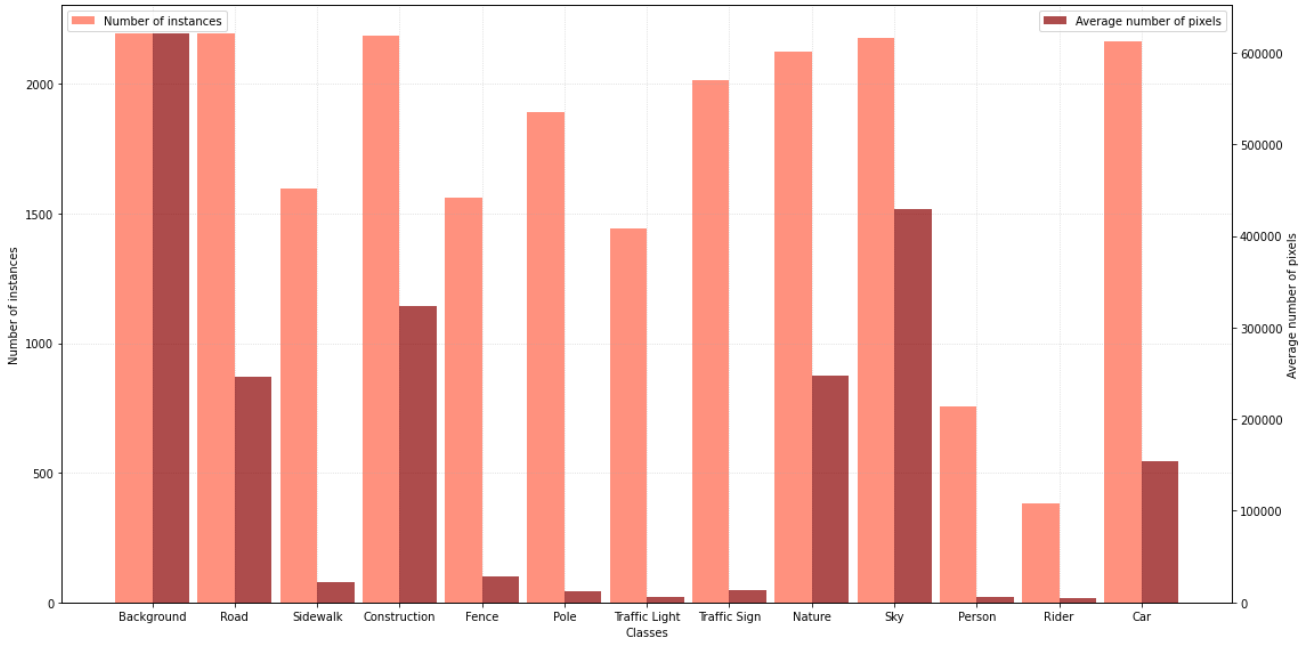
TRAIN_SOURCE_0000.png

TEST_0000.png

2. 탐색적 데이터 분석

범주 별 데이터 불균형 문제

- 모델 학습에 사용되는 학습 데이터에는 범주 별 불균형 문제가 존재함.
- 또한, 개수가 적은 범주들은 평균적으로 그 **픽셀의 개수**도 적음.
- 평가 지표 mIoU (mean Intersection over Union) 성능을 향상시키기 위해 소수 범주에 대한 다양한 크기의 예측이 필요함.
  -> Dice Loss 및 Multi-scale Test Time Augmentation(TTA)를 적용함.



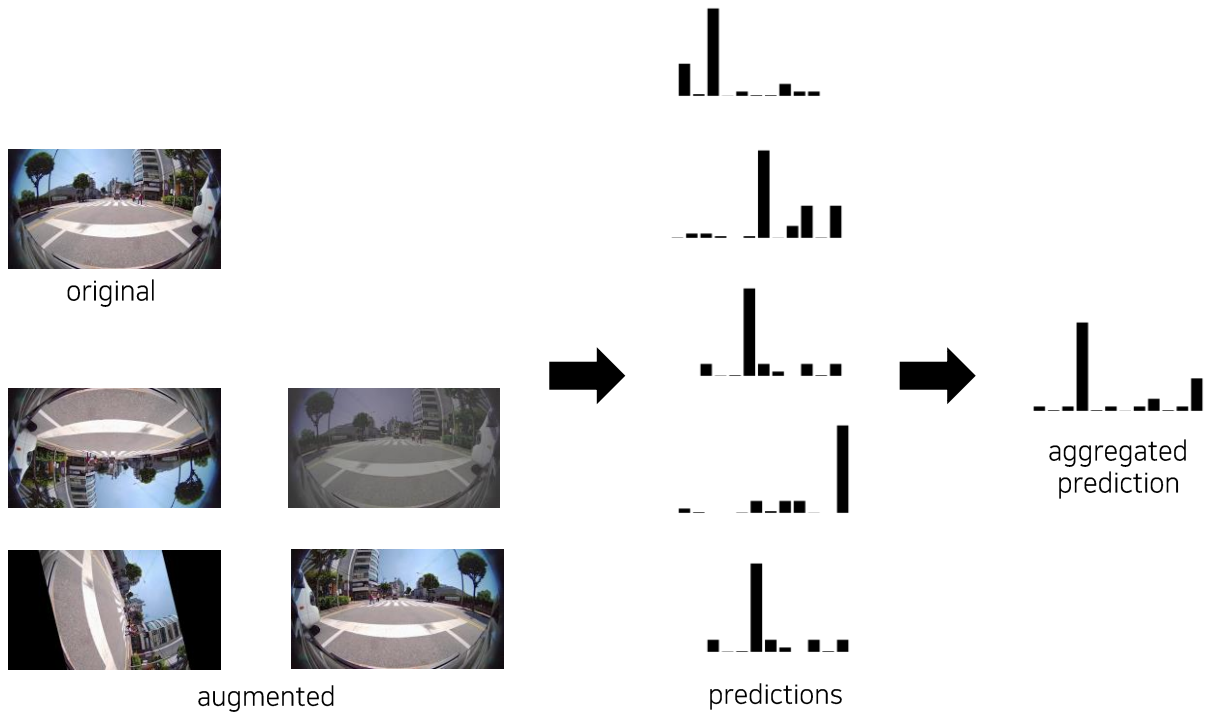
Rider class

2. 탐색적 데이터 분석

범주 별 데이터 불균형 문제

- Dice Loss란, ground truth에 해당하는 영역을 맞추는 것을 크게 평가하는 loss 함수로, 범주 별 데이터 불균형이 존재하는 semantic segmentation에서 자주 사용됨.
- Multi-scale Test Time Augmentation(TTA)는 inference 시 테스트 이미지 데이터에 augmentation을 적용하여 부족한 데이터셋의 문제점을 보완하는 기법임.

$$DiceLoss(y, \bar{p}) = 1 - \frac{2y\bar{p} + 1}{y + \bar{p} + 1}$$



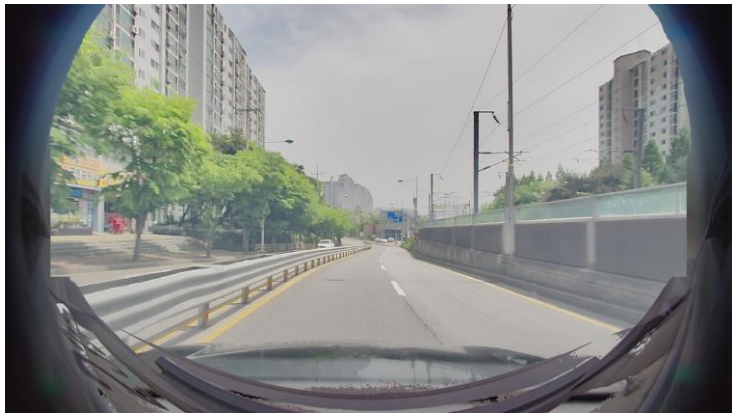
1) SUDRE, Carole H., et al. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer International Publishing, 2017. p. 240-248.

3. 방법론 : Fisheye Transformation

Fisheye Transformation for Domain Adaptation

- Fisheye 형태의 Target 이미지를 분할하기 위해 학습 데이터 이미지를 Fisheye Domain으로 변환함.
- Target 이미지 크기에 맞게 1920 x 1080 크기의 이미지로 resize함.
- Train target image를 활용하여 Fisheye data의 background를 정의함.
- 배경을 예측할 수 있도록 레이블을 다시 매핑함.
  - 0 ('Background'), 1~12 ('Road', 'Sidewalk', 'Construction', 'Fence', 'Pole', 'Traffic Light', 'Traffic Sign', 'Nature', 'Sky', 'Person', 'Rider', 'Car')

TRAIN_FISHEYE_0000.png



TRAIN_FISHEYE_0000.png

FISHEYE Ground Truth Visualization

3. 방법론 : Fisheye Transformation

Fisheye Transformation Pipeline

- Fisheye Transformation은 총 3가지 프로세스로 진행됨.

1. Train target image sample을 평균한 Mean target image를 생성함.

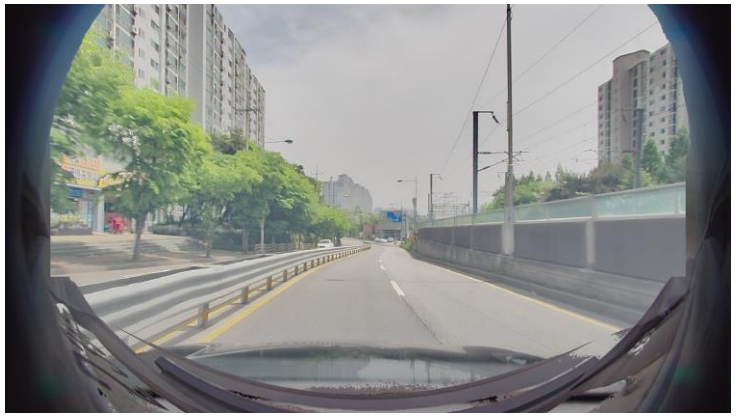
2. Mean target image를 채널 축을 기준으로 평균한 후 threshold(>108) 기준으로 bool type의 Mask target gt를 생성함.
   추가적으로, threshold로 제거되지 않은 noisy한 pixel에 대한 처리를 진행함. (이미지의 중앙 부분 [360:720, 120:1800])

3. Mask target gt를 기반으로 Mean target image를 source image와 결합한 Fisheye source image와 background class가 추가된 Fisheye source gt를 생성함.

Mean target image

Mask target gt

Fisheye source image

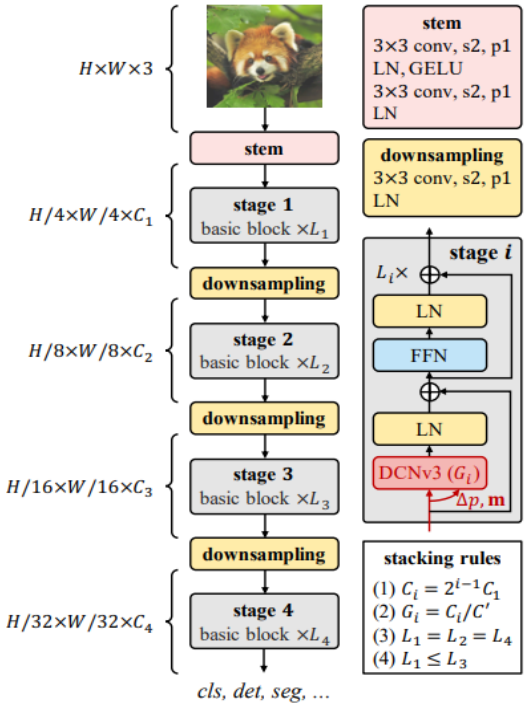


4. 방법론 : InternImage

InternImage Model

- InternImage는 Transformer의 Multi-head Self-attention 구조의 이점을 convolution에 반영한 Deformable Convolution v3(DCNv3)를 core operator로 사용하는 CNN 기반 모델임.
- Semantic Segmentation에서 좋은 성능을 나타내고 있는 모델들은 Vision Transformer(ViT) 기반 모델들이 지배적이지만, InternImage 모델은 Convolutional Neural Networks(CNN) 기반 모델임에도 불구하고 SOTA 성능을 나타내고 있음.

Overall Architecture of InternImage



Semantic segmentation performance on ADE20K

method	crop size	#params	#FLOPs	mIoU (SS)	mIoU (MS)
Swin-T [2]	512^2	60M	945G	44.5	45.8
ConvNeXt-T [21]	512^2	60M	939G	46.0	46.7
SLaK-T [29]	512^2	65M	936G	47.6	–
InternImage-T (ours)	512^2	59M	944G	47.9	48.1
Swin-S [2]	512^2	81M	1038G	47.6	49.5
ConvNeXt-S [21]	512^2	82M	1027G	48.7	49.6
SLaK-S [29]	512^2	91M	1028G	49.4	–
InternImage-S (ours)	512^2	80M	1017G	50.1	50.9
Swin-B [2]	512^2	121M	1188G	48.1	49.7
ConvNeXt-B [21]	512^2	122M	1170G	49.1	49.9
RepLKNet-31B [22]	512^2	112M	1170G	49.9	50.6
SLaK-B [29]	512^2	135M	1172G	50.2	–
InternImage-B (ours)	512^2	128M	1185G	50.8	51.3
Swin-L‡ [2]	640^2	234M	2468G	52.1	53.5
RepLKNet-31L‡ [22]	640^2	207M	2404G	52.4	52.7
ConvNeXt-L‡ [21]	640^2	235M	2458G	53.2	53.7
ConvNeXt-XL‡ [21]	640^2	391M	3335G	53.6	54.0
InternImage-L‡ (ours)	640^2	256M	2526G	53.9	54.1
InternImage-XL‡ (ours)	640^2	368M	3142G	55.0	55.3
SwinV2-G# [16]	896^2	3.00B	–	–	59.9
InternImage-H# (ours)	896^2	1.12B	3566G	59.9	60.3
BEiT-3# [17]	896^2	1.90B	–	–	62.8
FD-SwinV2-G# [26]	896^2	3.00B	–	–	61.4
InternImage-H# (ours) + Mask2Former [80]	896^2	1.31B	4635G	62.5	62.9

4. 방법론 : InternImage

Deformable Convolution v3

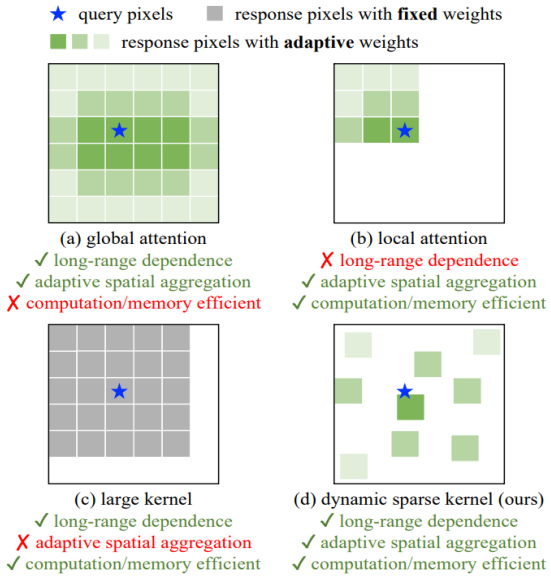
- DCNv2는 convolution neuron마다 독립적인 linear projection weights를 가지기 때문에 대규모 모델에서 효율이 떨어지는데, 이를 보완하기 위해 DCNv3에서는 가중치 w_k를 depth-wise와 point-wise로 분리하여 **뉴런 간의 가중치를 공유**하도록 함.
- 또한, Multi-head Self-attention의 multi-head와 유사하게 spatial aggregation process를 G 그룹으로 분할하여 **downstream task에 더 강력한 효과**를 냄.
- DCNv3 연산자를 사용함으로써 Long-Range Dependency와 Adaptive Spatial Aggregation를 도입해서 일반 convolution의 결함을 보완함.

DCNv2

$$\mathbf{y}(p_0) = \sum_{k=1}^{K} \mathbf{w}_k \mathbf{m}_k \mathbf{x}(p_0 + p_k + \Delta p_k),$$

DCNv3

$$\mathbf{y}(p_0) = \sum_{g=1}^{G} \sum_{k=1}^{K} \mathbf{w}_g \mathbf{m}_{gk} \mathbf{x}_g(p_0 + p_k + \Delta p_{gk}),$$



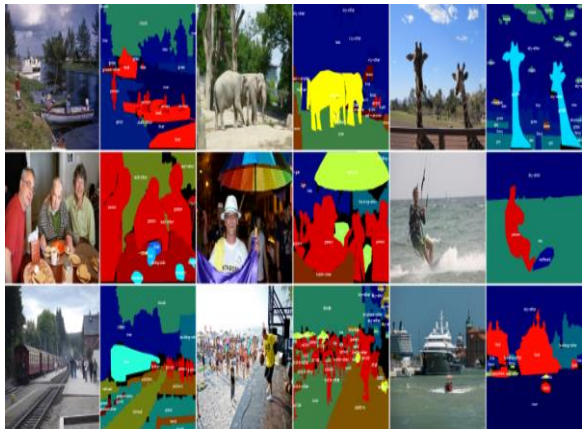
4. 방법론 : InternImage

InternImage Model

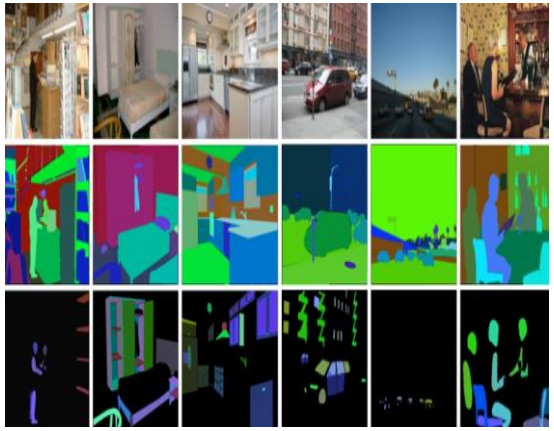
- InternImage 모델은 DCNv3를 사용하는 CNN 기반 모델이기 때문에, ViT 기반 모델에 비해 적은 데이터와 학습 시간으로도 효율적인 학습이 가능함.
  -> 한정된 학습 데이터만으로도 안정적으로 fine-tuning이 가능
- 제공받은 학습 데이터는 차량의 주변 상황을 담은 도로 이미지로 구성됨.
- 공개된 데이터셋인 ADE20K, Cityscapes, COCO-stuff는 대규모 데이터셋으로, 제공받은 train_source_image와 비슷한 domain의 이미지도 일부 포함하고 있음.
- 해당 데이터셋으로 사전 학습된 InternImage 모델(논문을 통해 공개된 사전학습 모델)을 이용함.

Example of Cityscapes dataset

Overview of COCO-stuff dataset



Overview of ADE20K dataset



1) https://www.cityscapes-dataset.com/, https://groups.csail.mit.edu/vision/datasets/ADE20K/, https://github.com/nightrome/cocostuff
2) https://github.com/OpenGVLab/InternImage/tree/master/segmentation/configs

5. 실험 결과

실험 설계

- InternImage-H 모델의 사전 훈련된 가중치는 아래와 같음.
  - UperNet: upernet_internimage_h_896_160k_ade20k.pth
  - Mask2Former: mask2former_internimage_h_896_80k_cocostuff2ade20k.pth
- Test Time Augmentation은 MultiScaleFlipAug를 사용함. (img_ratios = [0.8, 0.9, 1.0, 1.1, 1.2])
- 모든 실험은 최대한 동일한 조건으로 A100(40GB)과 A6000을 사용하여 학습함.
- InternImage 모델의 Hyperparameter는 다음과 같음.
  - Img_scale = (1920, 1080), crop_size = (960, 540)
  - Optimizer: AdamW, betas=(0.9, 0.999), weight_decay=0.05
  - Learning Rate: 0.00002, Polynomial Schedular
  - Random Seed: 826
- 평가 지표는 mIoU(mean Intersection over Union)로 각 class마다 Ground Truth와 Prediction의 교집합(Intersection = Area of Overlap)과 합집합(Area of Union)의 평균임.

5. 실험 결과

모델 성능

- 최종 제출 파일은 22000 Iter 값을 가진 InterImage-H + Fisheye Transform 모델임.
- 추가적으로 학습을 더 진행한 모델의 private mIoU 성능이 우수함을 확인함.
- Val mIoU는 background 범주를 제외한 mIoU 값임.

Model	Backbone	Private mIoU	Public mIoU	Val mIoU	#params	#FLOPs	Iter
InternImage-H + Fisheye Transform	UperNet	0.64483	0.6192	67.30	1.12B	3566G	20000
InternImage-H + Fisheye Transform	Mask2Former	0.66456	0.62271	66.01	1.31B	4635G	9000
InternImage-H + Fisheye Transform +TTA	UperNet	0.65114	0.62775	68.03	1.12B	3566G	14000
InternImage-H + Fisheye Transform + TTA	Mask2Former	**0.66886**	**0.63133**	70.34	1.31B	4635G	22000
InternImage-H + Fisheye Transform + TTA	Mask2Former	**0.67288**	**0.62905**	-	1.31B	4635G	40000

6. 적용 가능성 및 결론

- 본 경진대회에서는 Fisheye Transformation을 활용하는 InternImage 기반의 이미지 분할 모델을 제안함.
- Domain Adaptation 방법론을 적용하는 Network는 최적화가 어렵고 실험적으로 성능이 떨어짐을 확인함.
- Fisheye Transformation은 Fisheye target domain에 적용하기 쉬우며 성능을 쉽게 향상시킬 수 있는 장점을 가짐.
- InternImage는 현재 Semantic Segmentation 분야에서 state-of-the-art의 성능을 갖는 모델임.
- ADE20K, Cityscape 등의 자율주행 관련 데이터로 사전 학습된 가중치를 사용하여 쉽게 우수한 성능의 모델을 만들 수 있음.
- Pseudo Labeling 및 Soft Voting Ensemble을 활용하여 성능을 향상 시킬 수 있을 것으로 기대됨.
- 하지만, 해당 방법론을 사용할 시 실제 적용 가능성에서 비용이 매우 많이 드는 문제가 발생함.
- 단일 모델로도 충분히 우수한 성능을 갖는 모델을 만들 수 있음을 실험적으로 확인함.
- Test Time Augmentation을 적용 시 추론 시간이 매우 오래 걸리는 한계점을 가짐.
- Fisheye가 아닌 다른 Target Domain에 대한 이미지 분할 시 추가적인 알고리즘을 개발해야 하는 한계점을 가짐.
- 재현 가능한 코드를 공개함. (https://github.com/GNOEYHEAT/Fisheye-InternImage)

Any Questions?

taehyeong93@korea.ac.kr