2025 하계 세미나

Robust Foundation Model

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
양진철

Outline

- Background

- Papers
  - RobustSAM: Segment Anything Robustly on Degraded Images [CVPR 2024]
  - Depth Anything at Any Condition [arXiv 2025]

Background

- What is Foundation Model?
  - 대규모 dataset으로 학습하여 다양한 이미지에 대해서 일반화가 가능한 모델
  - Segment Anything Model (SAM)
    - SA-1B dataset (11.1M images와 1B가 넘는 masks로 구성)
  - Depth Anything v2
    - DA-2K dataset (595K synthetic images 와 62M pseudo-labeled real images로 구성)

		SA-1B	
	# countries	#imgs	#masks
Africa	54	300k	28M
Asia & Oceania	70	3.9M	423M
Europe	47	5.4M	540M
Latin America & Carib.	42	380k	36M
North America	4	830k	80M
high income countries	81	5.8M	598M
middle income countries	108	4.9M	499M
low income countries	28	100k	9.4M

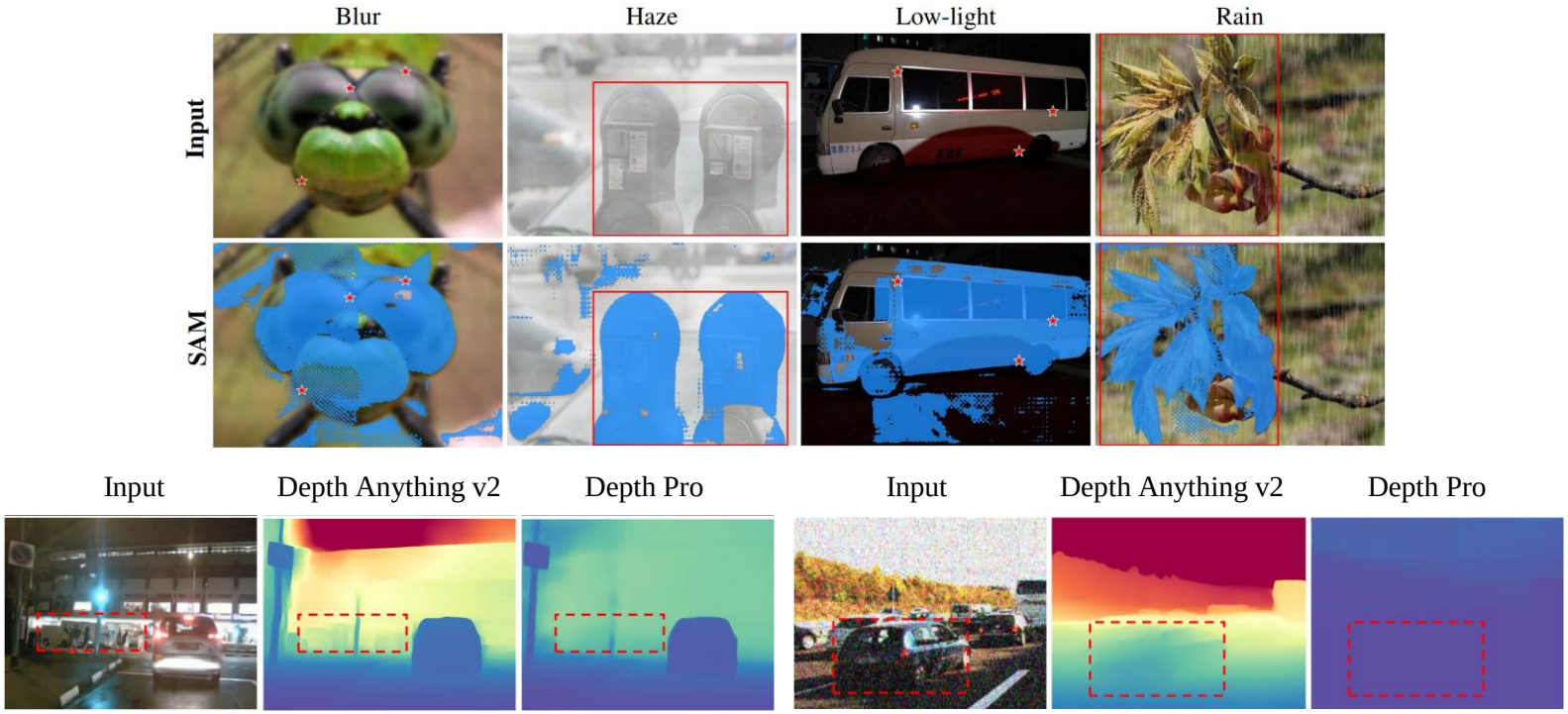
< SA-1B dataset >

Dataset	Indoor	Outdoor	# Images
Precise *Synthetic* Images (595K)			
BlendedMVS [92]	✓	✓	115K
Hypersim [58]	✓		60K
IRS [77]	✓		103K
TartanAir [79]	✓	✓	306K
VKITTI 2 [9]		✓	20K
Pseudo-labeled *Real* Images (62M)			
BDD100K [97]		✓	8.2M
Google Landmarks [81]		✓	4.1M
ImageNet-21K [60]	✓	✓	13.1M
LSUN [98]	✓		9.8M
Objects365 [65]	✓	✓	1.7M
Open Images V7 [35]	✓	✓	7.8M
Places365 [103]	✓	✓	6.5M
SA-1B [33]	✓	✓	11.1M

< DA-2K dataset >

Background

- Robustness and Scalability when Challenging Scenarios
  - Degraded conditions
    - Low lighting, noise, blur, adverse weather, and compression artifacts
    - Degraded conditions에서는 일반화 성능 하락



< Results of foundation model under different degradations using unseen datasets >

RobustSAM: Segment Anything Robustly on Degraded Images [CVPR 2024]

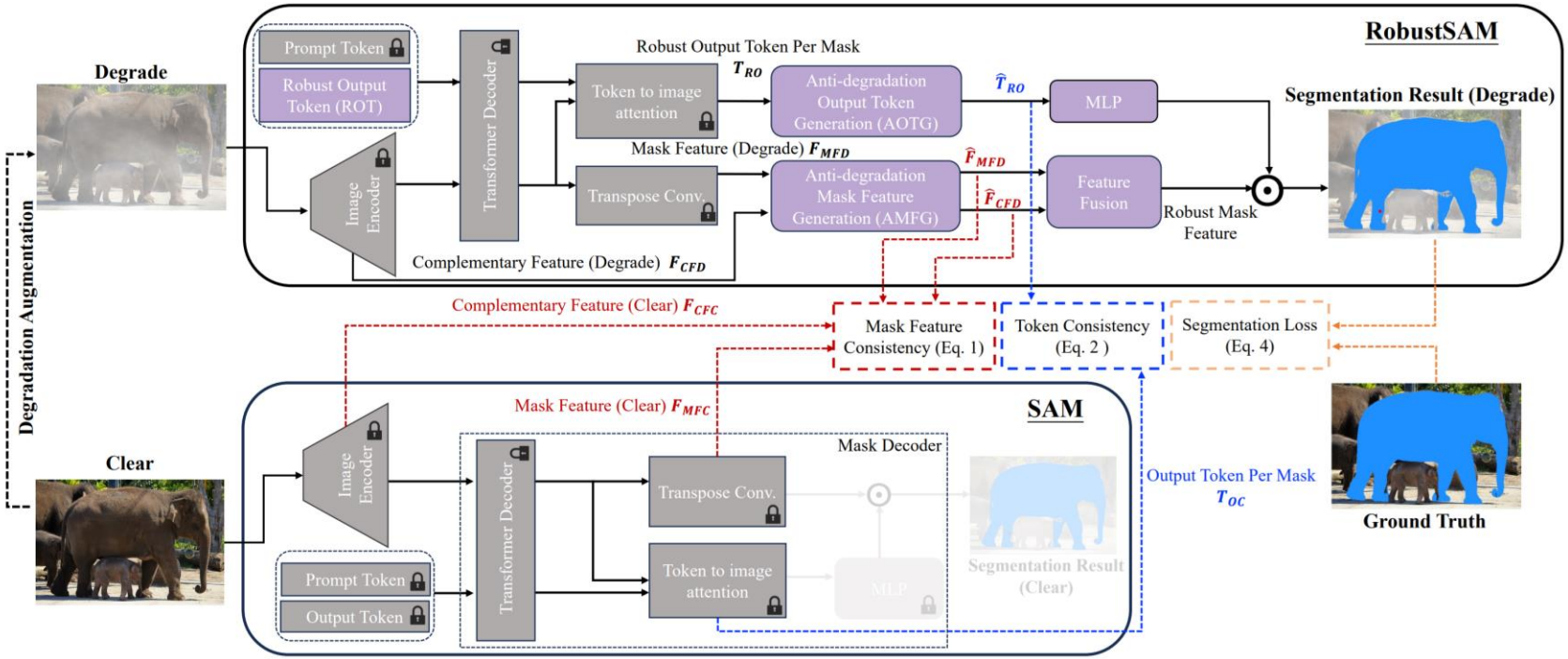
1) Chen, Wei-Ting, et al. "Robustsam: Segment anything robustly on degraded images." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

RobustSAM[1)]

- Keyword
  - SAM, Image degradation, Dataset
- Introduction
  - 다양한 image degradation의 segmentation을 robust하기 위한 RobustSAM 제안
  - 다양한 degradations를 포함하는 Robust-Seg dataset 제안
- Analysis
  - Image 전처리 후 segmentation → Image quality ↑, Segmentation performance ?
  - Degraded image로 SAM 자체를 fine-tuning → catastrophic forgetting

RobustSAM[1)]

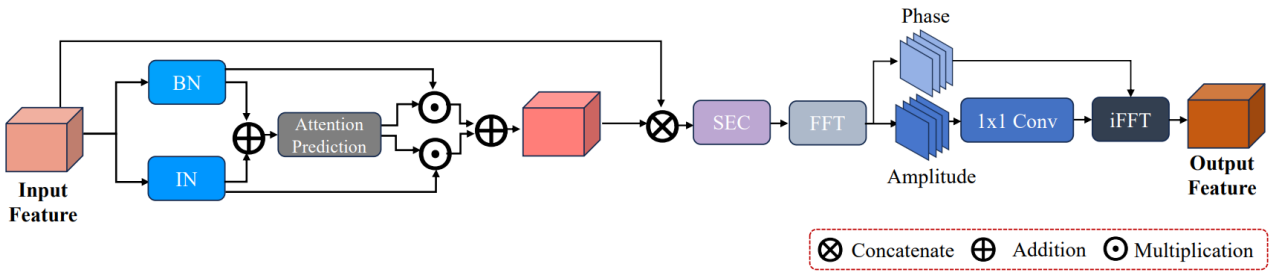
- Overview of RobustSAM
  - RobustSAM
    - SAM
    - AntiDegradation Output Token Generation (AOTG)
    - AntiDegradation Mask Feature Generation (AMFG)



< Overview of RobustSAM >

RobustSAM[1)]

- Proposed Method
  - AntiDegradation Mask Feature Generation (AMFG)
    - Degradation으로 인한 손상 영역을 마스크로 식별하고, 해당 영역의 특징을 보완 및 복원하는 특화된 feature를 생성해 모델의 복원 성능을 향상시키는 역할
      - Complementary Feature (F_{CFD}), mask Feature (F_{MFD}) 에 적용
    - 구조
      - Instance Normalization (IN)
        - ✓ 물체의 형태 정보 유지, degrdation으로 변질될 수 있는 색·밝기·질감·노이즈 같은 요소 제거
      - Batch Normalization (BN)
        - ✓ IN에서 발생할 수 있는 세부 정보 손실을 보완



< Overview of AMFG >

RobustSAM[1)]

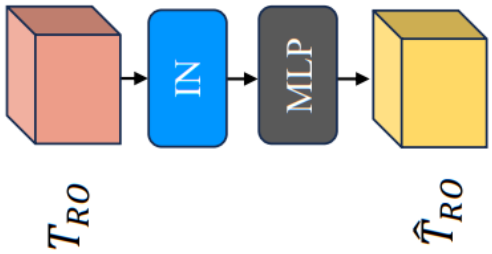
- Proposed Method
  - AntiDegradation Mask Feature Generation (AMFG)
    - 구조
      - Squeeze-and-Excitation (SEC)
        - ✓ 채널별 중요도를 조절해, 중요한 채널은 강화하고 덜 중요한 채널은 약화
      - Fourier Degradation Suppression
        - ✓ 공간영역에서 주파수영역으로 변환하기 위해 Fourier transform 사용
        - ✓ Amplitude 성분을 활용해 image degradation과 관련된 정보 포착
          - 1×1 컨볼루션을 적용하여 열화 요소를 분리·제거하는 데 집중하며,
        - ✓ Phase 성분은 구조·경계·형태 정보를 담아 객체의 모양 보존
    - Mask Feature Consistency Loss
      - $\mathcal{L}_{\text{MFC}}$: Refiend feature가 clear image를 input으로 했을 때의 feature와 일관성을 유지하기 위한 loss

$$\mathcal{L}_{\text{MFC}} = \left|\left|\hat{F}_{CFD} - F_{CFC}\right|\right|_2 + \left|\left|\hat{F}_{MFD} - F_{MFC}\right|\right|_2$$

RobustSAM[1)]

- Proposed Method
  - AntiDegradation Output Token Generation (AOTG)
    - 각 마스크별 Robust Output Token (T_{RO})에서 열화와 관련된 정보를 제거하여 정제
      - T_{RO}는 주로 경계의 명확성을 보장하는 기능을 수행하여 텍스처 정보는 적음
        - ✓ Degradation에 sensitive 정보를 걸러내기 위해 가벼운 모듈 사용
        - ✓ 2층의 IN 과 단일 MLP 사용
    - Token Consistency Loss
      - $\mathcal{L}_{\text{TC}}$: Refiend token이 clear image를 input으로 했을 때의 token과 일관성을 유지하도록 보장하기 위한 loss

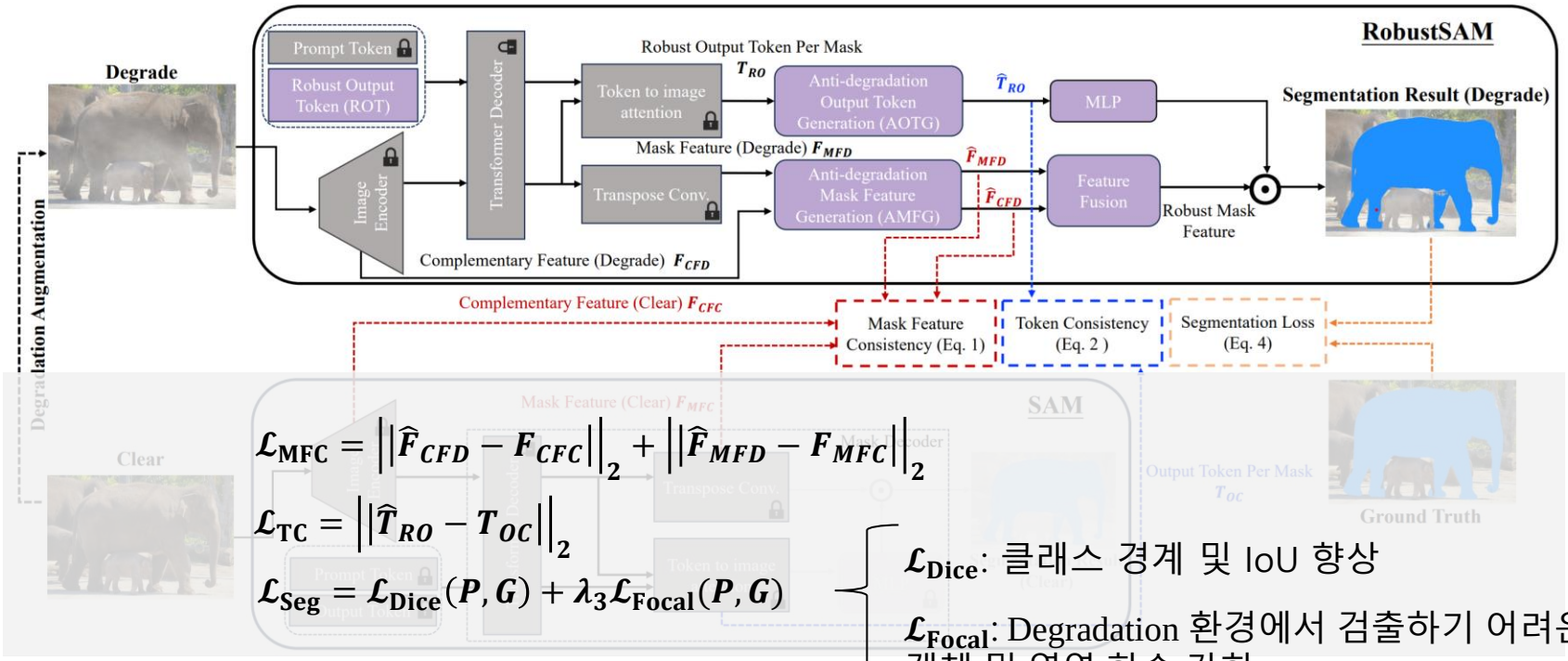
$$\mathcal{L}_{\text{TC}} = \left|\left|\hat{T}_{RO} - T_{OC}\right|\right|_2$$



< Overview of AOTG >

RobustSAM[1)]

- Proposed Method
  - Overall Loss



$$\mathcal{L}_{\text{MFC}} = \left|\left|\hat{F}_{CFD} - F_{CFC}\right|\right|_2 + \left|\left|\hat{F}_{MFD} - F_{MFC}\right|\right|_2$$

$$\mathcal{L}_{\text{TC}} = \left|\left|\hat{T}_{RO} - T_{OC}\right|\right|_2$$

$$\mathcal{L}_{\text{Seg}} = \mathcal{L}_{\text{Dice}}(P, G) + \lambda_3 \mathcal{L}_{\text{Focal}}(P, G)$$

$$\mathcal{L}_{\text{Overall}} = \mathcal{L}_{MFC} + \lambda_1 \mathcal{L}_{TC} + \lambda_2 \mathcal{L}_{\text{Seg}}$$

$\mathcal{L}_{\text{Dice}}$: 클래스 경계 및 IoU 향상

$\mathcal{L}_{\text{Focal}}$: Degradation 환경에서 검출하기 어려운 객체 및 영역 학습 강화

P, G: Prediction, GT mask

RobustSAM[1)]

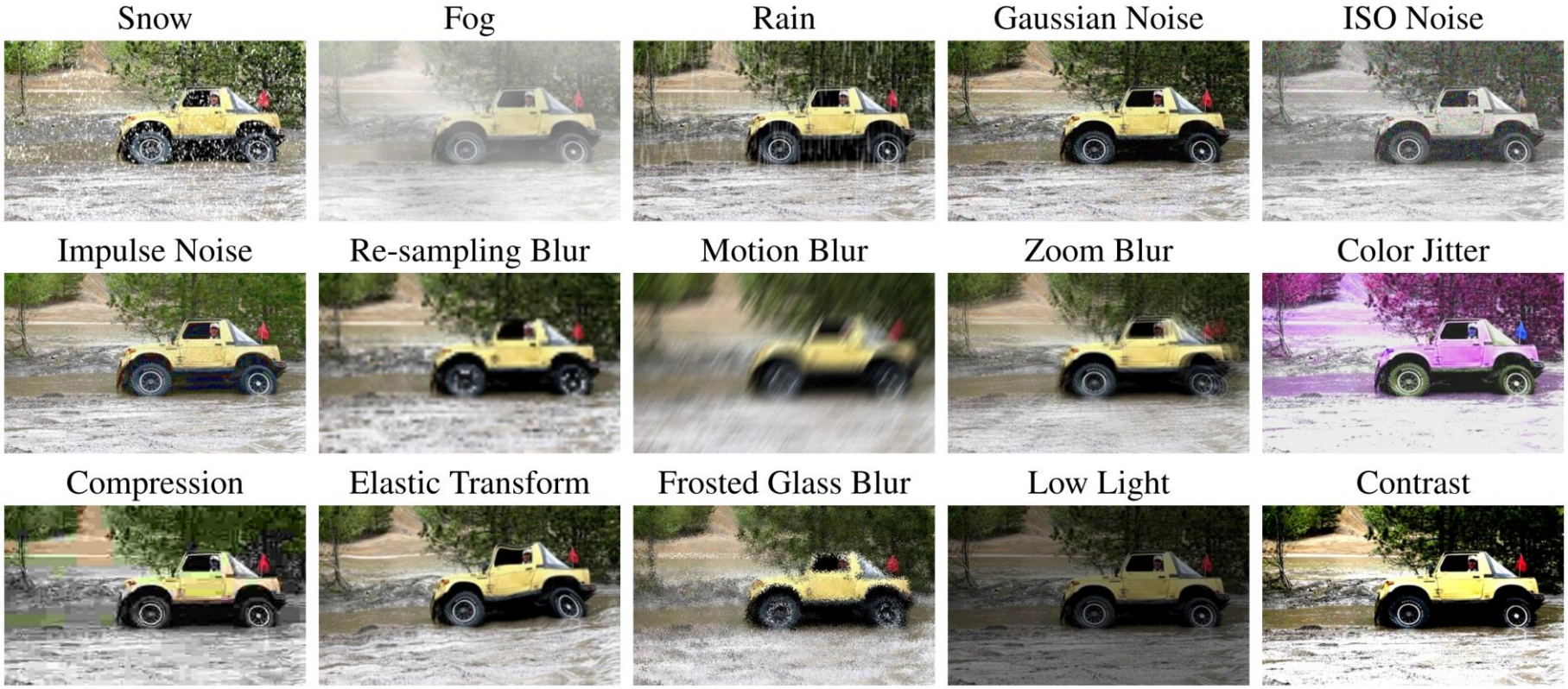
- Robust-Seg dataset
  - 총 688,000 image-mask pairs로 구성
    - Clear image: 43,000 장
      - LVIS, ThinObject-5k, MSRA10K, NDD20, STREETS, FSS-1000, COCO 로 구성
    - Degradation image: 645,000 장
      - Clear image에 15가지 타입으로 synthetic degradation augmentation 진행
  - Training sets
    - MSRA10K, ThinObject-5k, and LVIS
  - Test sets
    - MSRA10k and LVIS
  - For zero-shot generalization
    - NDD20, STREETS, FSS-1000, and COCO / BDD-100k and LIS (Real-world degradations)

	LVIS [19]	ThinObject-5k [48]	MSRA10K [11]	NDD20 [71]	STREETS [67]	FSS-1000 [45]	COCO [51]	**Total**
Image Number	20252	4748	10000	1000	1000	1000	5000	43000

< Data composition of Robust-Seg dataset >

RobustSAM[1)]

- Robust-Seg dataset
  - Synthetic degradation augmentation
    - Blur, noise, low light, adverse weather conditions – 15 types



< Images with synthetic degradations >

RobustSAM[1)]

- Experimental Results
  - Quantitative Results

Method	Degrade		Clear		Average	
	IoU	PA	IoU	PA	IoU	PA
SAM	0.8194	0.9108	0.8402	0.9235	0.8207	0.9116
HQ-SAM	0.8358	0.9202	0.8604	0.9328	0.8373	0.9210
AirNet+SAM	0.8157	0.9193	0.8236	0.9294	0.8162	0.9199
URIE+SAM	0.8217	0.9125	0.8450	0.9245	0.8231	0.9132
RobustSAM	**0.8609**	**0.9640**	**0.8726**	**0.9649**	**0.8616**	**0.9641**

< Performance Comparison on test set of MSRA10k >

Method	Performance Metrics			
	AP	AP_S	AP_M	AP_L
SAM	0.5002	0.3168	0.4292	0.5243
HQ-SAM	0.5052	0.2920	0.4267	0.5517
AirNet+SAM	0.4926	0.3068	0.4263	0.5187
URIE+SAM	0.4980	0.3186	0.4319	0.5215
RobustSAM	**0.5130**	**0.3192**	**0.4416**	**0.5518**

< Zero-shot segmentation comparison on the whole COCO >

Method	Degrade		Clear		Average	
	IoU	PA	IoU	PA	IoU	PA
SAM	0.7341	0.9181	0.7415	0.9282	0.7346	0.9187
HQ-SAM	0.7405	0.9242	0.7502	0.9319	0.7411	0.9246
AirNet+SAM	0.7352	0.9198	0.7419	0.9293	0.7356	0.9204
URIE+SAM	0.7336	0.9182	0.7406	0.9277	0.7340	0.9188
RobustSAM	**0.7506**	**0.9327**	**0.7592**	**0.9339**	**0.7511**	**0.9328**

< Performance Comparison on test set of LVIS >

Method	Point		Bounding Box	
	IoU	Dice	IoU	Dice
SAM	0.3056	0.3837	0.8808	0.9171
HQ-SAM	0.2943	0.3712	0.8877	0.9245
AirNet+SAM	0.3245	0.4550	0.8776	0.9129
URIE+SAM	0.3042	0.3828	0.8799	0.9165
RobustSAM	**0.3717**	**0.8926**	**0.8958**	**0.9416**

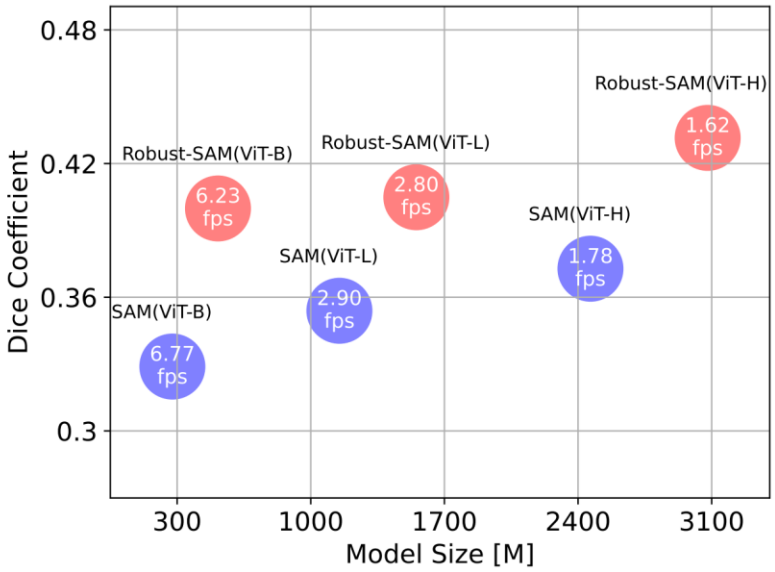
< Zero-shot segmentation comparison on the whole BDD-100k and LIS >

RobustSAM[1)]

- Ablation Studies
  - 제안한 module의 효과
  - SAM과 RobustSAM 비교
    - Performance, speed, and model size

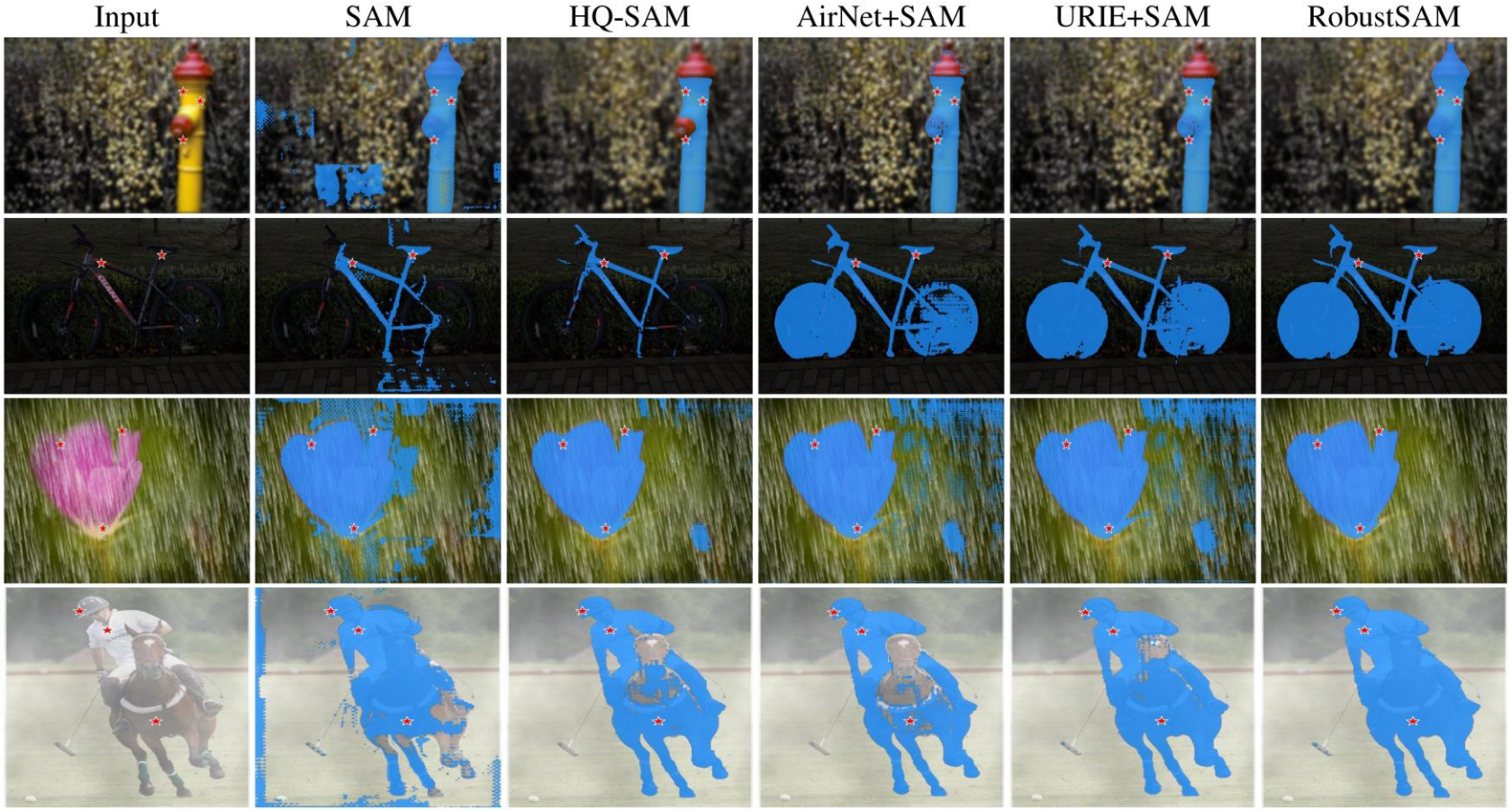
Module	Metric	
	IoU	PA
Baseline		
SAM	0.3056	0.8911
SAM-Finetune	0.1871	0.7691
SAM-Finetune Decoder	0.2476	0.8691
SAM-Finetune Output Token	0.3194	0.9036
RobustSAM		
w AMFG	0.3455	0.9059
w AMFG-F	0.3535	0.9120
w AMFG-F+AOTG	0.3651	0.9193
w AMFG-F+AOTG+ROT (**ALL**)	**0.3717**	**0.9226**

< Efficacy of Proposed Modules >



RobustSAM[1)]

- Experimental Results
  - Qualitative Results



< Qualitative Analysis of Segmentation >

Depth Anything at Any Condition [arXiv 2025]

1) Sun, Boyuan, et al. "Depth Anything at Any Condition." *arXiv preprint arXiv:2507.01634* (2025).

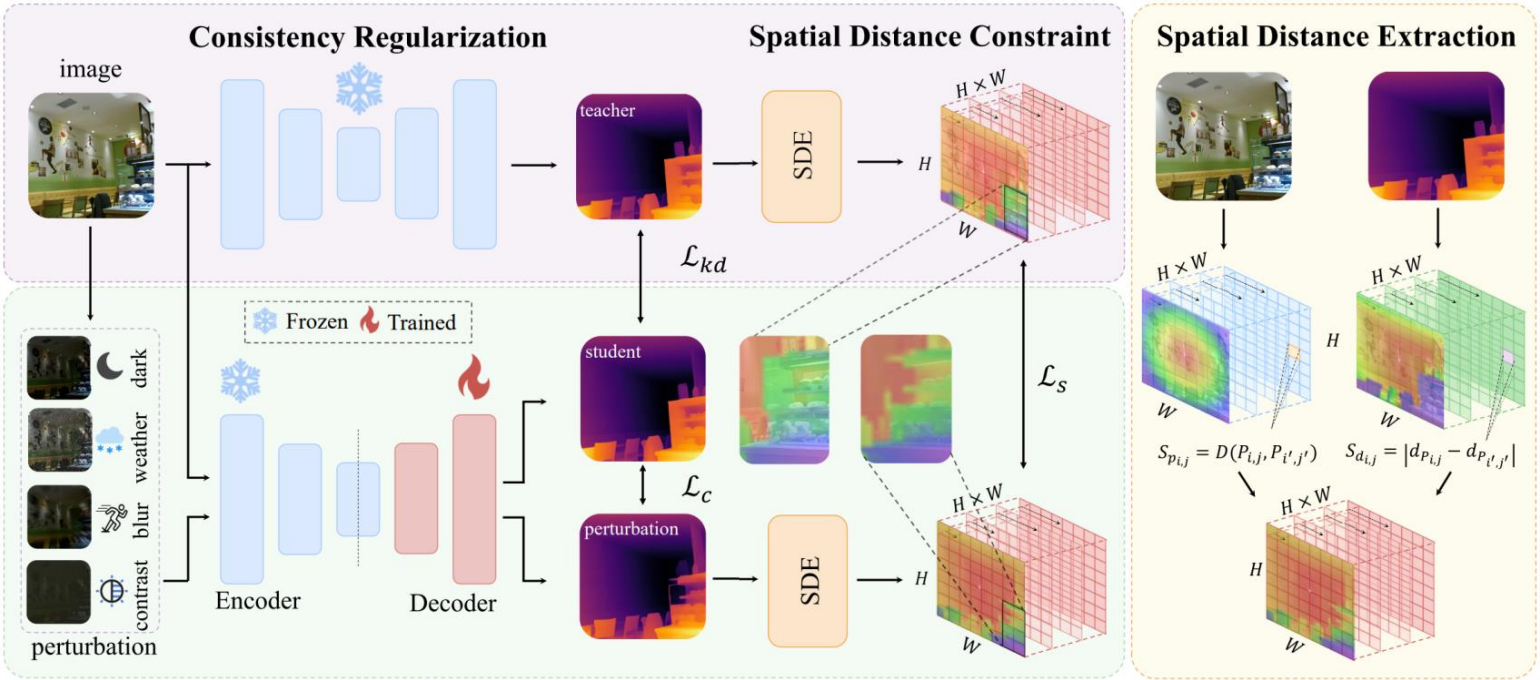
Depth Anything at Any Condition[1)]

- Keyword
  - Monocular depth estimation, Image degradation, Data augmentation, Knowledge distillation
- Introduction
  - 다양한 image degradation의 MDE 강건성을 강화한 zero-shot 모델 DepthAnything-AC 제안
- Analysis
  - Image-GT depth 쌍이 부족
  - Degraded image로 Depth Anything 자체를 fine-tuning → catastrophic forgetting

< Histogram of post-Softmax activations >

Depth Anything at Any Condition[1)]

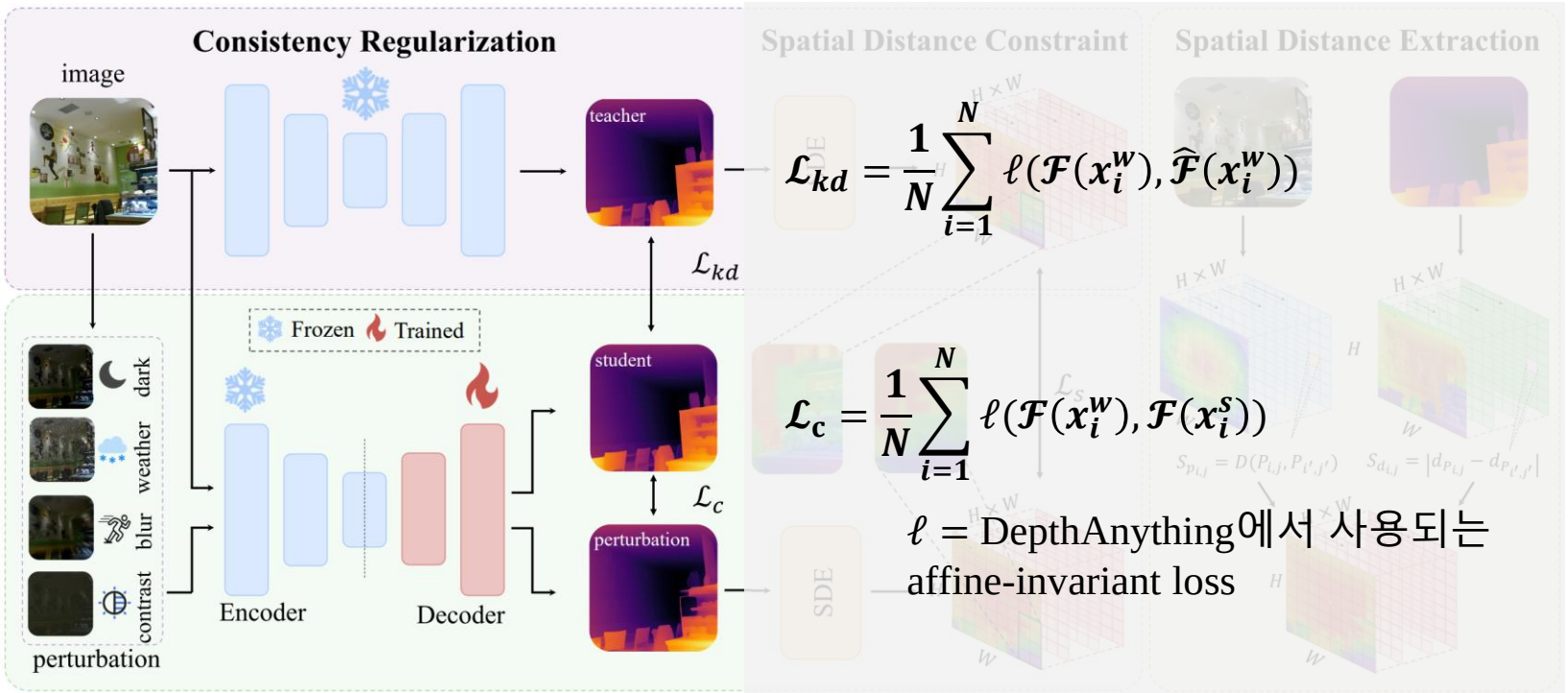
- Overview of DepthAnything-AC
  - DepthAnything-AC
    - Depth Anything
    - Perturbation-Based Consistency
    - Spatial Distance Constraint



< Training pipeline for DepthAnyting-AC >

Depth Anything at Any Condition[1)]

- Proposed Method
  - Perturbation-Based Consistency Framework
    - 복잡한 시나리오에서 기존 MDE 모델의 성능을 확장하기 위한 framework



Depth Anything at Any Condition[1)]

- Proposed Method
  - Spatial Distance Constraint
    - 객체 경계를 정확히 구분하고 세밀한 디테일을 포착하기 위한 방법
      - 이미지 평면에서 패치 간의 relative positional distance : S_p
      - Disparity distance : $S_{d_{i,j}}$

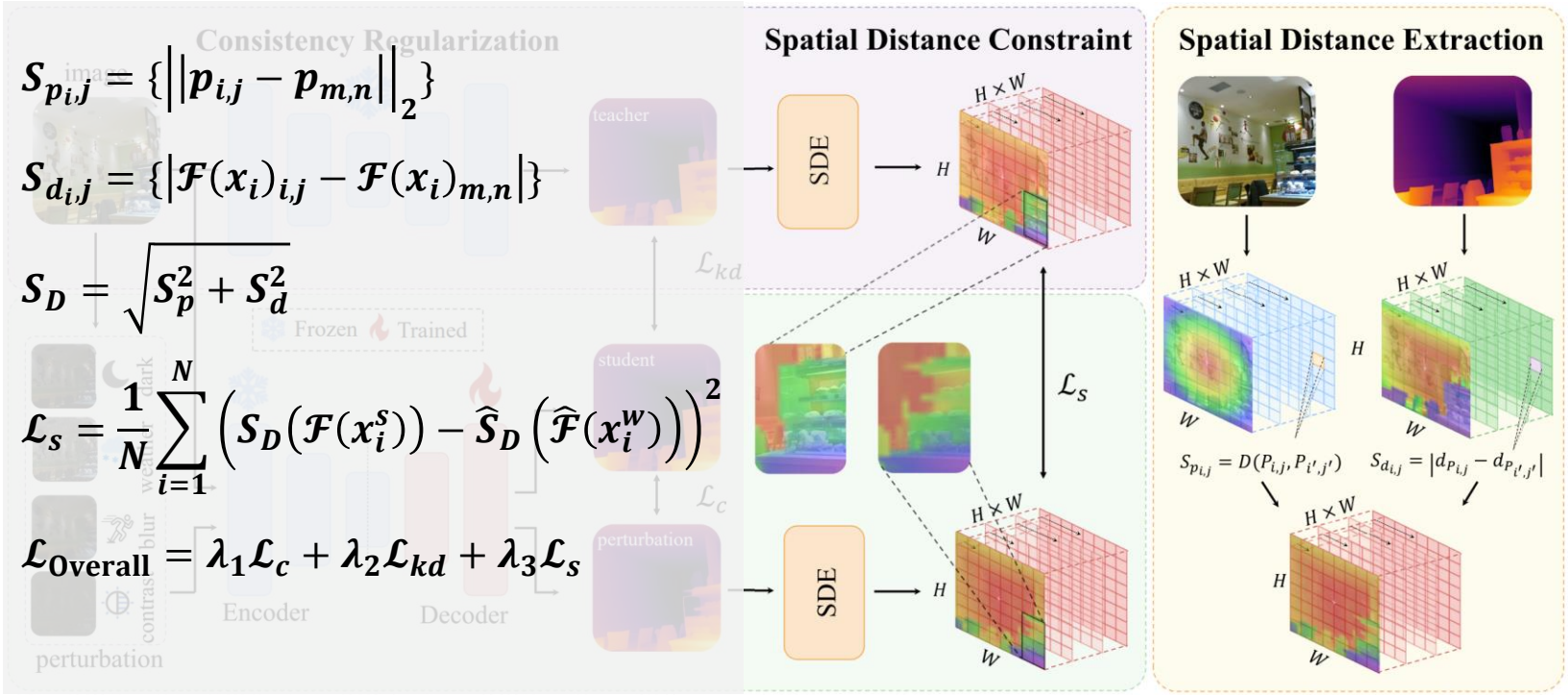
$$S_{p_{i,j}} = \{\left|\left|p_{i,j} - p_{m,n}\right|\right|_2\}$$

$$S_{d_{i,j}} = \{\left|\mathcal{F}(x_i)_{i,j} - \mathcal{F}(x_i)_{m,n}\right|\}$$

$$S_D = \sqrt{S_p^2 + S_d^2}$$

$$\mathcal{L}_s = \frac{1}{N}\sum_{i=1}^{N}\left(S_D\left(\mathcal{F}(x_i^s)\right) - \widehat{S}_D\left(\widehat{\mathcal{F}}(x_i^w)\right)\right)^2$$

$$\mathcal{L}_{\text{Overall}} = \lambda_1\mathcal{L}_c + \lambda_2\mathcal{L}_{kd} + \lambda_3\mathcal{L}_s$$



Depth Anything at Any Condition[1)]

- DepthAnything-AC dataset
  - Fine-tuning을 위한 dataset
    - 총 540,000 장으로 DA-2K dataset의 1% 미만
  - Test sets (For zero-shot generalization)
    - DA-2K, KITTI-C (synthetic benchmark)
    - NuScenes-night, Robotcar-night, Driving-Stereo (real-world degradations)
    - KITTI, NYU-D, Sintel, ETH3D, DIODE (general benchmarks)

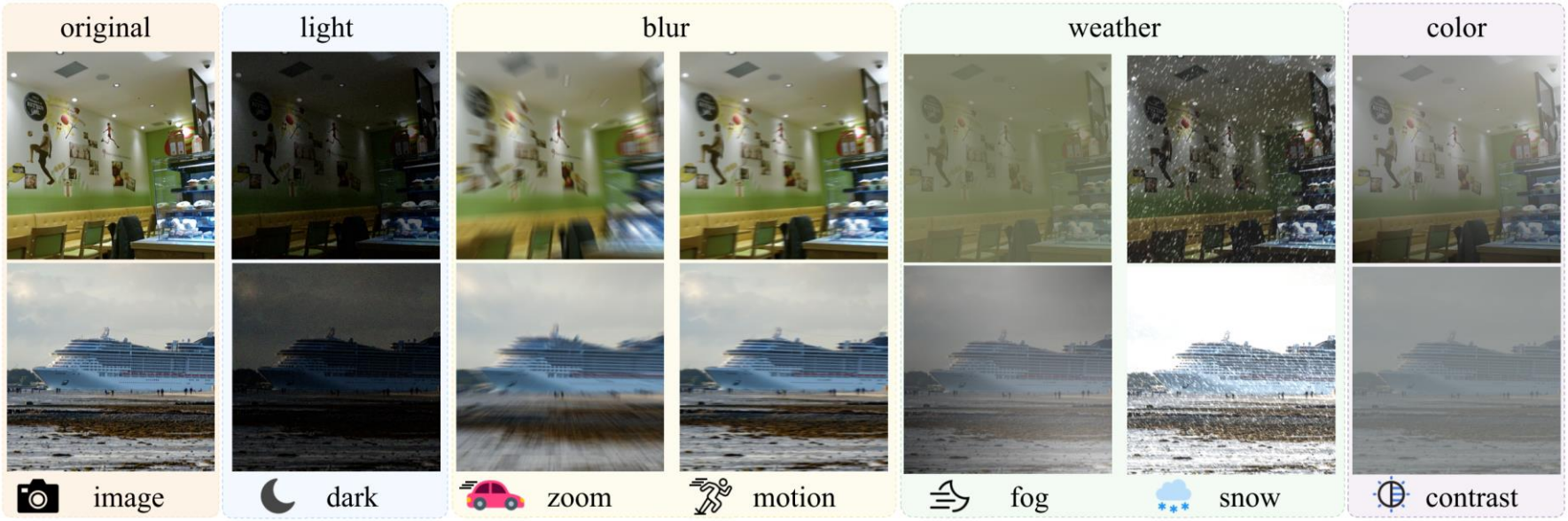
Dataset	Indoor	Outdoor	Samples	Used	Ratio
ADE20k [105, 106]	✓	✓	20K	20K	100%
MegaDepth [44]		✓	128K	100K	78%
DIML [13, 37, 12, 38, 36]	✓	✓	927K	80K	8.6%
VKITTI2 [9]		✓	42K	42K	100%
HRWSI [22, 85]		✓	20K	20K	100%
SA-1B [39]	✓	✓	11.1M	140K	1.3%
COCO [47]	✓	✓	120K	120K	100%
Pascal VOC 2012 [17]	✓	✓	10K	10K	100%
AODRaw [45]	✓	✓	80K	8K	10%

< Datasets used for training >

1) Sun, Boyuan, et al. "Depth Anything at Any Condition." *arXiv preprint arXiv:2507.01634* (2025).

Depth Anything at Any Condition[1)]

- DepthAnything-AC dataset
  - Synthetic degradation augmentation
    - Darkness, weather (fog, snow), blur (motion, zoom), and contrast perturbations
    - Darkenss는 항상 적용
    - Blur와 weather는 각각 확률로 0.1, 0.2 적용



< Visualization of perturbation types >

Depth Anything at Any Condition[1)]

- Experimental Results
  - Quantitative Results

Method	Encoder	DA-2K	DA-2K dark	DA-2K fog	DA-2K snow	DA-2K blur
DynaDepth [99]	ResNet	0.655	0.652	0.613	0.605	0.633
EC-Depth [67]	ViT-S	0.753	0.732	0.724	0.713	0.701
STEPS [104]	ResNet	0.577	0.587	0.581	0.561	0.577
robustdepth [61]	ViT-S	0.724	0.716	0686	0.668	0.680
weather-depth [78]	ViT-S	0.745	0.724	0.716	0.697	0.666
DepthPro[7]	ViT-S	0.947	0.872	0.902	0.793	0.772
DepthAnything V1 [89]	ViT-S	0.884	0.859	0.836	0.880	0.821
DepthAnything V2 [90]	ViT-S	0.952	0.910	0.922	0.880	0.862
DepthAnything-AC	ViT-S	**0.953**	**0.923**	**0.929**	**0.892**	**0.880**

< Quantitative results on the enhanced multi-condition DA-2K benchmark >

Method	Encoder	NuScenes-night		RobotCar-night		DS-rain		DS-cloud		DS-fog	
		AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
DynaDepth [99]	ResNet	0.381	0.394	0.512	0.294	0.239	0.606	0.172	0.608	0.144	0.901
EC-Depth [67]	ViT-S	0.243	0.623	0.228	0.552	0.155	0.766	0.158	0.767	0.109	0.861
STEPS [104]	ResNet	0.252	0.588	0.350	0.367	0.301	0.480	0.252	0.588	0.216	0.641
robustdepth [61]	ViT-S	0.260	0.597	0.311	0.521	0.167	0.755	0.168	0.775	0.105	0.882
weather-depth [78]	ViT-S	-	-	-	-	0.158	0.764	0.160	0.767	0.105	0.879
Syn2Real [87]	ViT-S	-	-	-	-	0.171	0.729	-	-	0.128	0.845
DepthPro [7]	ViT-S	0.218	0.669	0.237	0.534	**0.124**	**0.841**	0.158	0.779	**0.102**	**0.892**
DepthAnything V1 [89]	ViT-S	0.232	0.679	0.239	0.518	0.133	0.819	0.150	**0.801**	0.098	0.891
DepthAnything V2 [90]	ViT-S	0.200	0.725	0.239	0.518	0.125	0.840	0.151	0.798	0.103	0.890
DepthAnything-AC	ViT-S	**0.198**	**0.727**	**0.227**	**0.555**	0.125	0.840	**0.149**	**0.801**	0.103	0.889

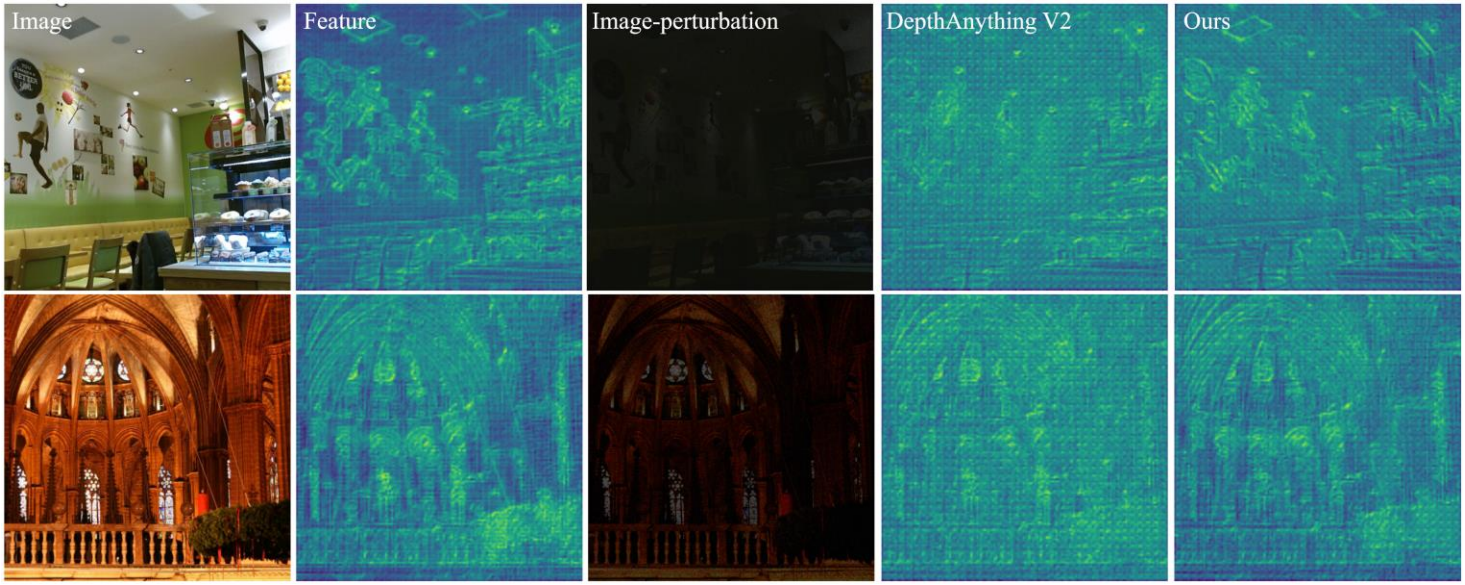
< Zero-shot relative depth estimation results on real complex benchmarks >

Depth Anything at Any Condition[1)]

- Experimental Results
  - Ablation studies

Method	NuScenes-night		Robotcar-night		DS-cloud		Snow		Gaussian	
	AbsRel↓	$\delta 1$↑	AbsRel↓	$\delta 1$↑	AbsRel↓	$\delta 1$↑	AbsRel↓	$\delta 1$↑	AbsRel↓	$\delta 1$↑
Unfrozen	0.218	0.691	0.248	0.494	0.151	0.798	0.114	0.872	0.155	0.789
Frozen	**0.198**	**0.727**	**0.227**	**0.555**	**0.149**	**0.801**	**0.114**	**0.873**	**0.153**	**0.793**

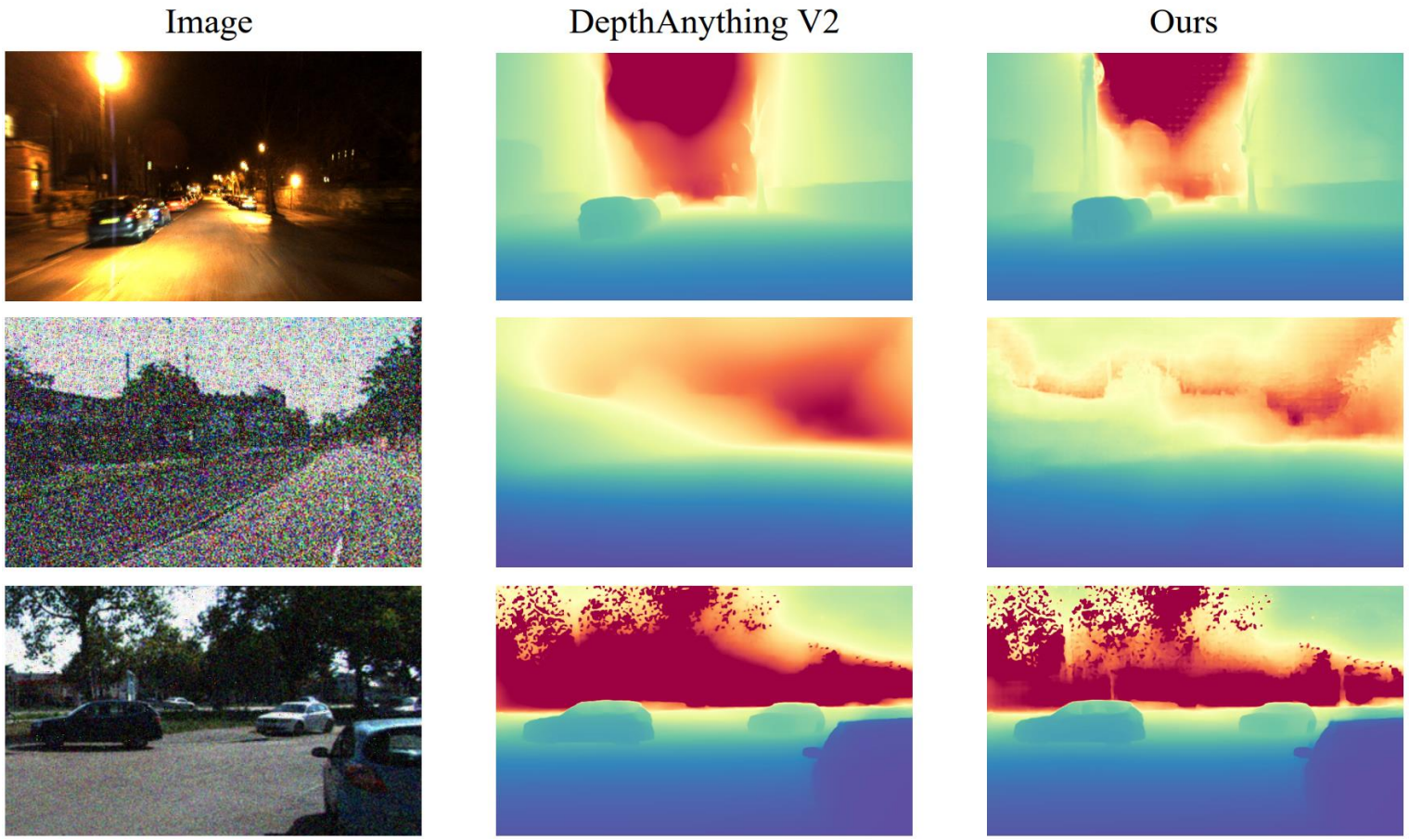
< The importance of maintaining the frozen state of the encoder >



< Visualization of feature representations >

Depth Anything at Any Condition[1)]

- Experimental Results
  - Qualitative Results



Thank you