Paper Survey

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

Outline

- Language-driven All-in-one Adverse Weather Removal
  - CVPR 2024
- HazeCLIP: Towards Language Guided Real-World Image Dehazing
  - arXiv 2024

Language-driven All-in-one Adverse Weather Removal

Hao Yang[1], Liyuan Pan[1], Yan Yang[2], and Wei Liang[1]
[1]Beijing Institute of Technology [2] The Australian National University
{hao.yang, liyuan.pan, liangwei}@bit.edu.cn, {yan.yang}@anu.edu.au

CVPR 2024

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Introduction

- Leverage the power of **pre-trained vision-language (PVL)** models to enrich the diversity of weather-specific knowledge.

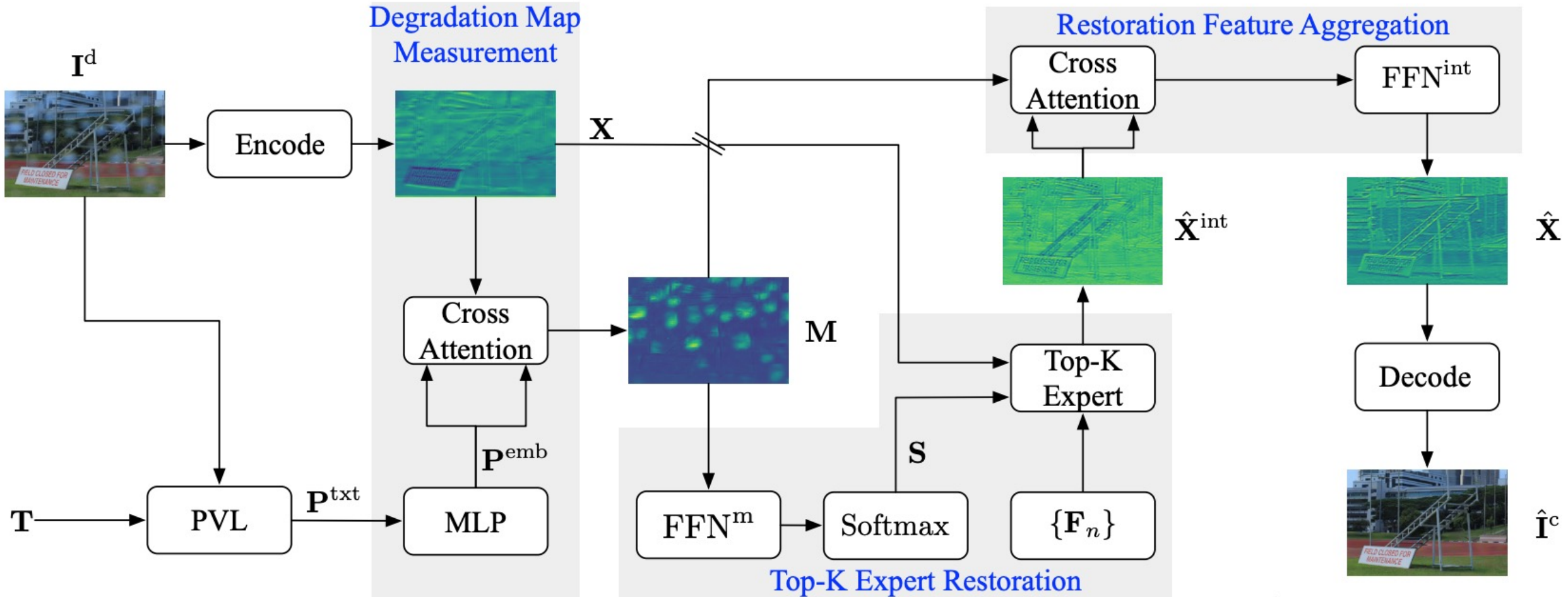
- With the guidance of **degradation prior**, author sparsely select restoration experts from a candidate list and further improve result in another module.

- LDR framework adaptively learn the weather-specific and shared knowledge to handle various weather conditions (e.g., unknown or mixed weather).

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Framework


Degradation Map Measurement
Restoration Feature Aggregation
$\mathbf{I}^{\mathrm{d}}$
Encode
$\mathbf{X}$
Cross Attention
FFN$^{\mathrm{int}}$
$\hat{\mathbf{X}}^{\mathrm{int}}$
$\hat{\mathbf{X}}$
Cross Attention
$\mathbf{M}$
Top-K Expert
Decode
$\mathbf{P}^{\mathrm{emb}}$
$\mathbf{S}$
$\mathbf{T}$
PVL
$\mathbf{P}^{\mathrm{txt}}$
MLP
FFN$^{\mathrm{m}}$
Softmax
$\{\mathbf{F}_n\}$
$\hat{\mathbf{I}}^{\mathrm{c}}$
Top-K Expert Restoration

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

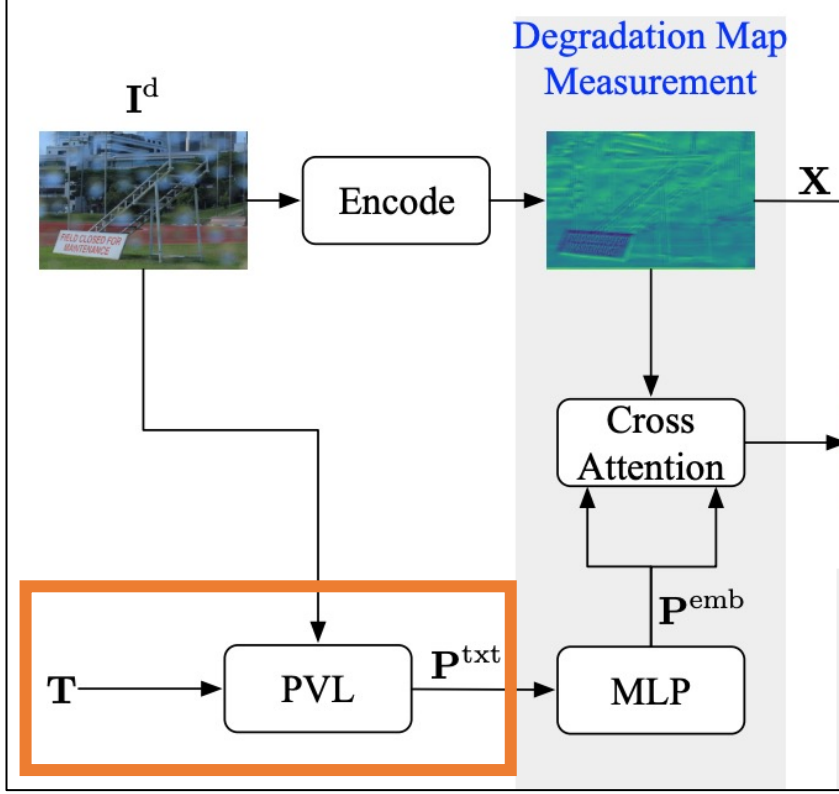
Examples text descriptions

Please describe about the weather in the picture.

The weather in the picture is rainy, as evidenced by the presence of raindrop on the cars and the overall atmosphere.

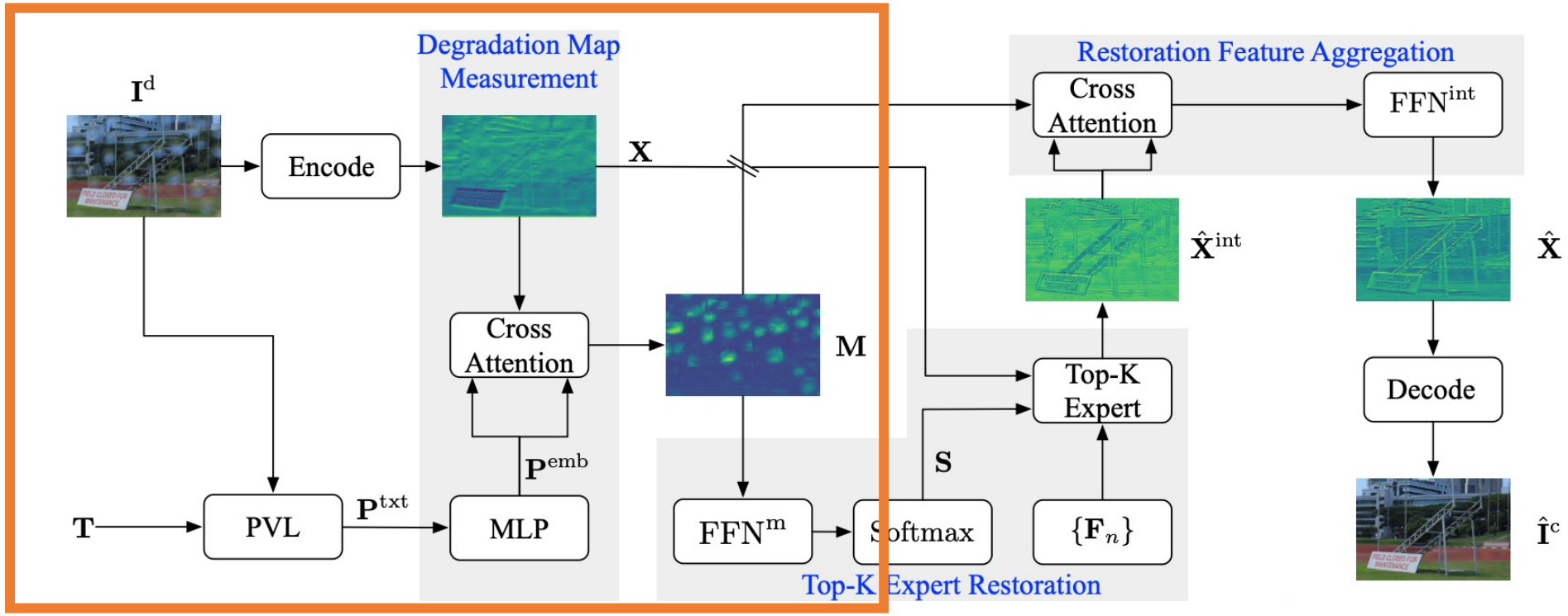
Please describe the type of weather, intensity, and obscured areas in the picture.

The weather in the image is snowy, the intensity of the snowfall is high. The area covered by the snow includes a residential neighborhood, with houses in the scene are obscured by snowflakes.



$$\mathbf{P}^{\mathrm{txt}} = \mathbf{VL}(\mathbf{I}^{\mathrm{c}}, \mathbf{T}), \quad \mathbf{P}^{\mathrm{txt}} \in \mathbb{R}^{L \times C^{\mathrm{vl}}}, \qquad (1)$$

Degradation Map Measurement



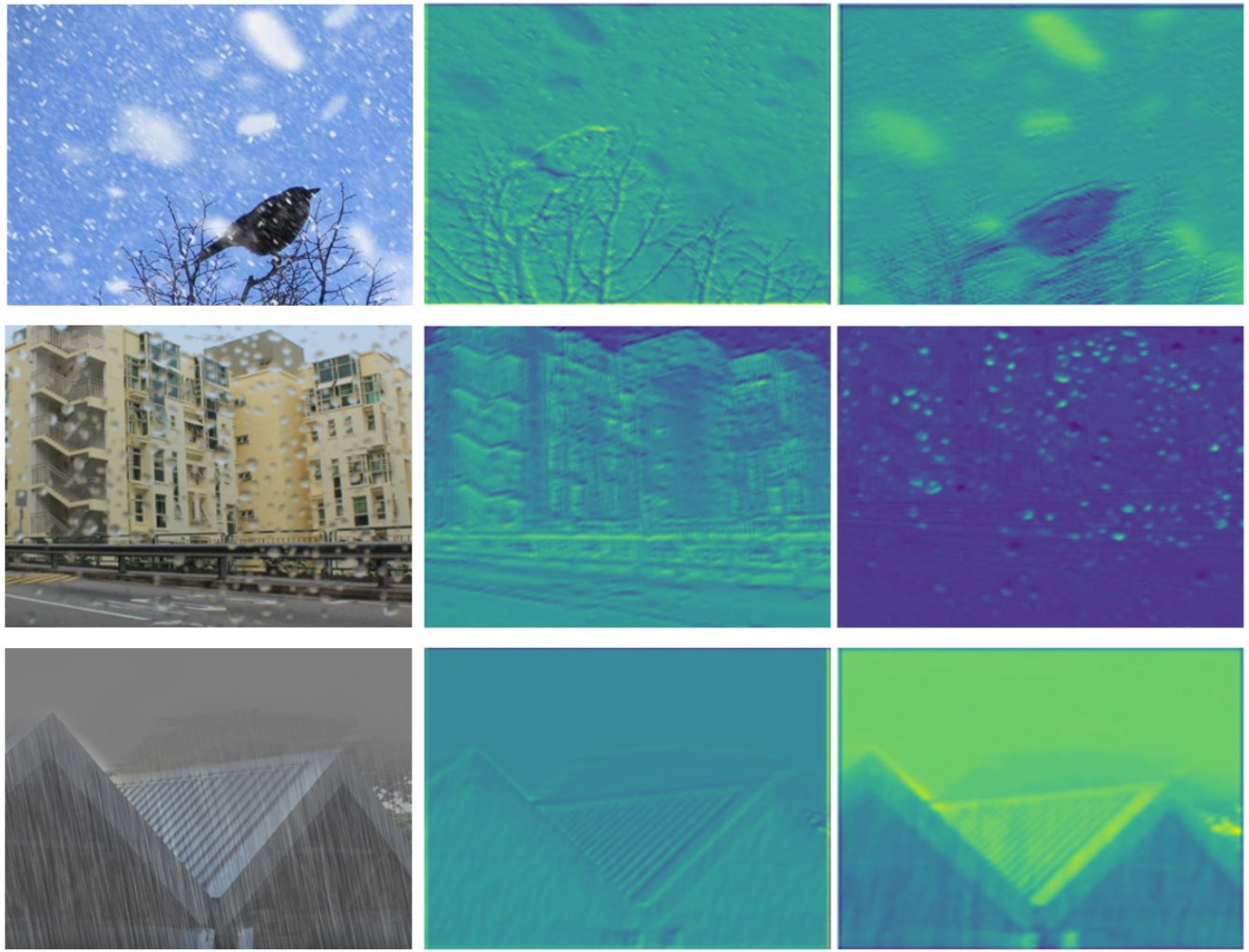
$$\mathbf{P}^{\mathrm{txt}} = \mathrm{VL}(\mathbf{I}^{\mathrm{c}}, \mathbf{T}) , \quad \mathbf{P}^{\mathrm{txt}} \in \mathbb{R}^{L \times C^{\mathrm{vl}}} , \qquad (1)$$

$$\mathbf{P}^{\mathrm{emb}} = \mathrm{MLP}(\mathbf{P}^{\mathrm{txt}}) , \quad \mathbf{P}^{\mathrm{emb}} \in \mathbb{R}^{L \times C} . \qquad (2)$$

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{\mathrm{q}_1} , \quad \mathbf{K} = \mathbf{P}^{\mathrm{emb}}\mathbf{W}^{\mathrm{k}_1} , \quad \mathbf{V} = \mathbf{P}^{\mathrm{emb}}\mathbf{W}^{\mathrm{v}_1} , \quad (3)$$

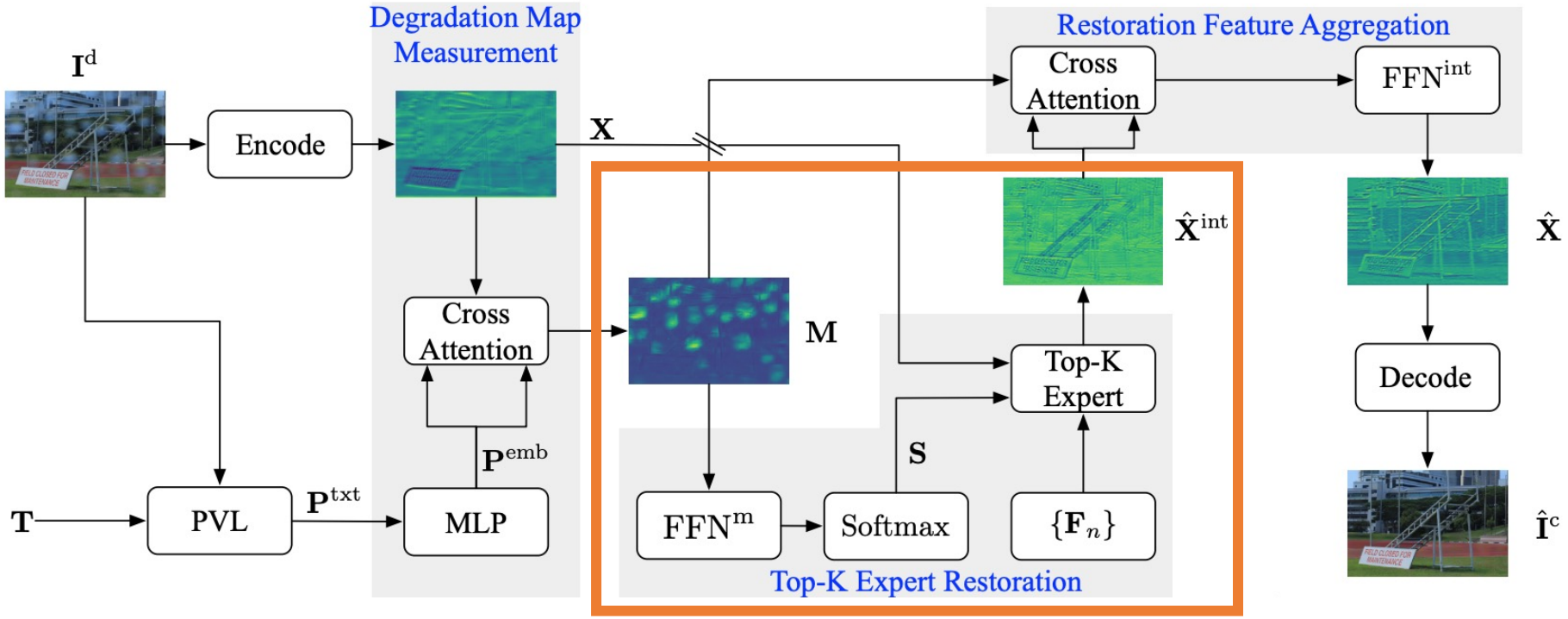
$$\mathbf{M} = \mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\mathbf{Q}\mathbf{K}^{\top}\right)\mathbf{V} , \qquad (4)$$

Visualization



(a) $\mathbf{I}^{\mathrm{d}}$ (b) $\mathbf{X}$ (c) $\mathbf{M}$

Top-K Expert Restoration



degradation map $\mathbf{M} \in \mathbb{R}^{H \times W \times C}$

$$\mathbf{S} = \mathrm{Softmax}(\mathrm{FFN}^{\mathrm{m}}(\mathbf{M}))$$

pixel-wise selection score $\mathbf{S} \in \mathbb{R}^{H \times W \times N}$

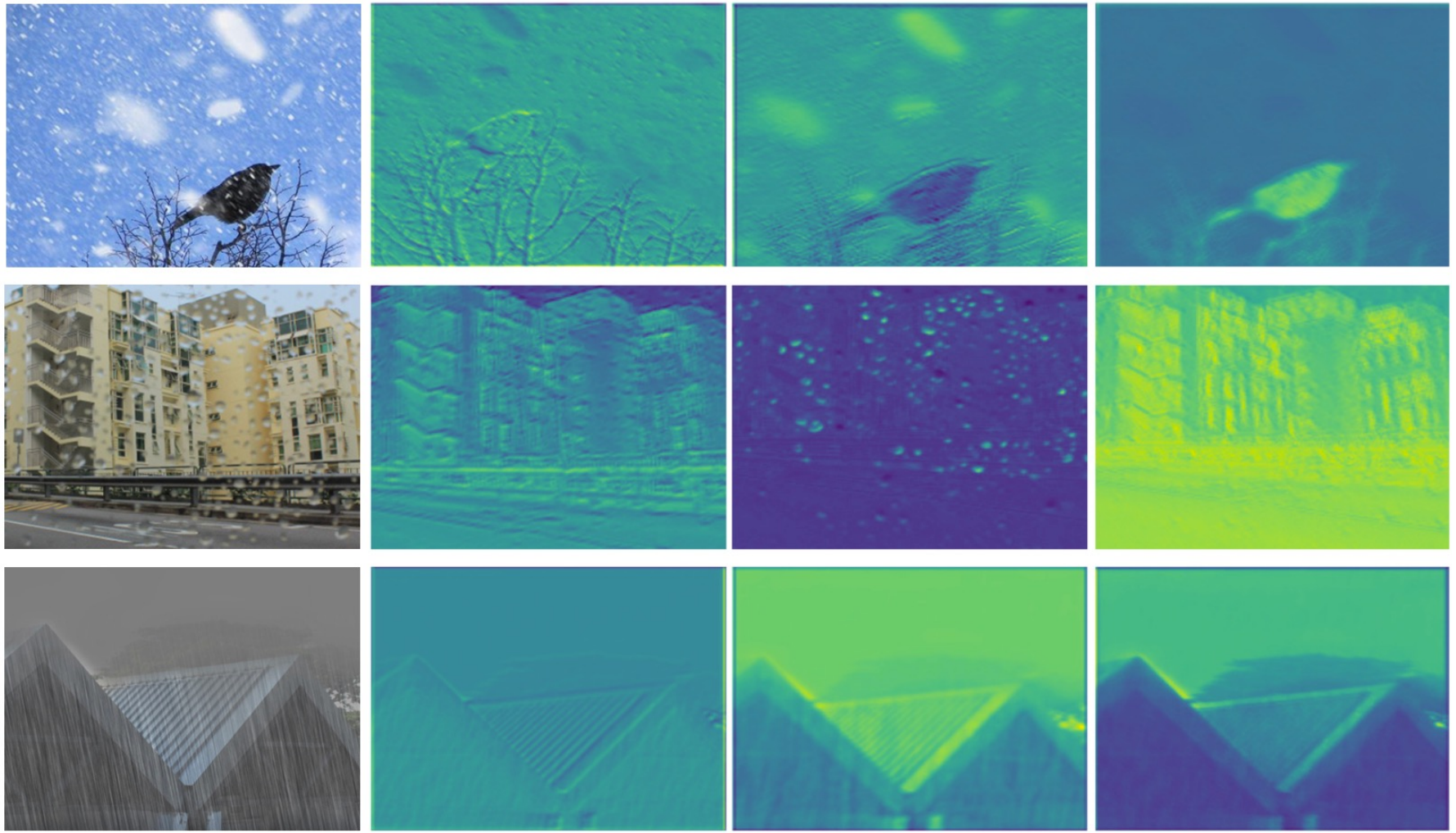
$$\{\mathbf{F}_n | n = 1, ..., N\}$$

$$\hat{\mathbf{X}}^{\mathrm{int}}(i,j) = \sum_{k=1}^{K} \mathbf{S}(i,j,\rho(k)) \cdot \mathcal{E}(i,j,\rho(k)) \,, \tag{5}$$

$$\mathcal{E}(i,j,\rho(k)) = \sum_{\Delta i, \Delta j} \mathbf{X}(i+\Delta i, j+\Delta j) \cdot \mathbf{F}_{\rho(k)}(\Delta i, \Delta j) \,,$$

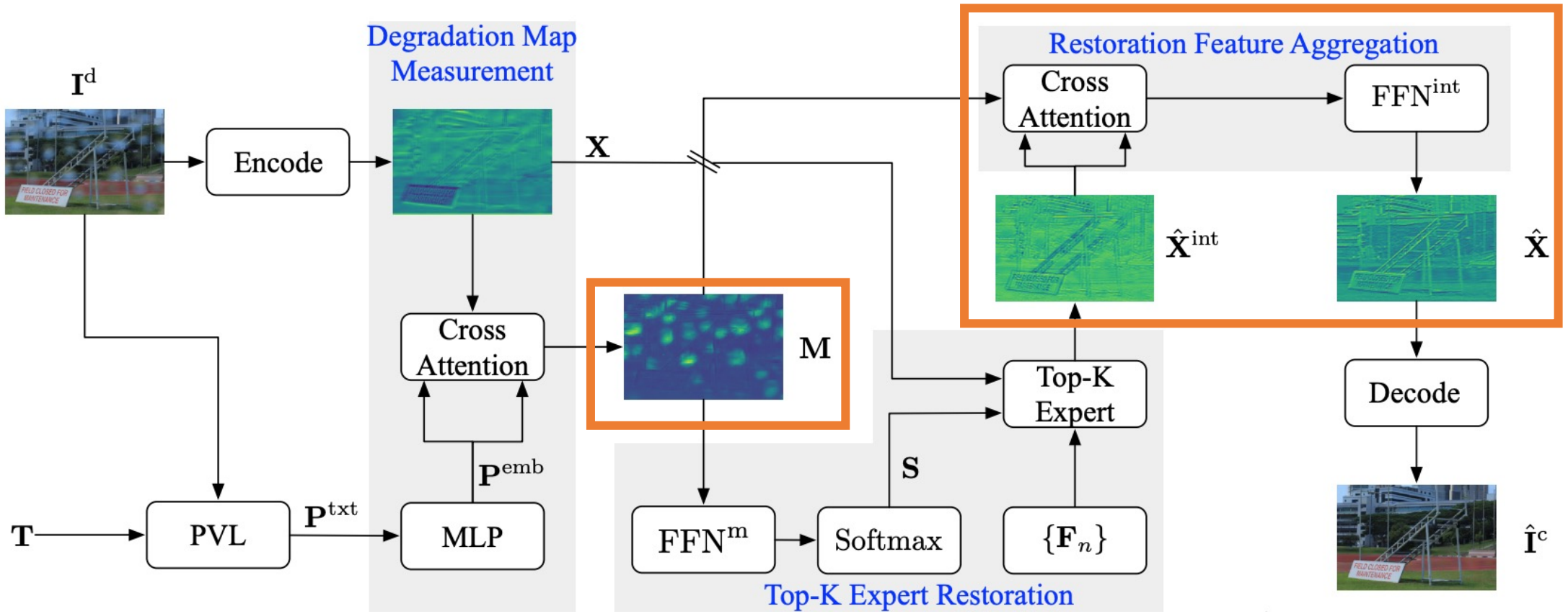
where $\rho(k)$ is the index of the selected k-th expert.

Visualization



(a) $\mathbf{I}^{\mathrm{d}}$ (b) $\mathbf{X}$ (c) $\mathbf{M}$ (d) $\hat{\mathbf{X}}^{\mathrm{int}}$

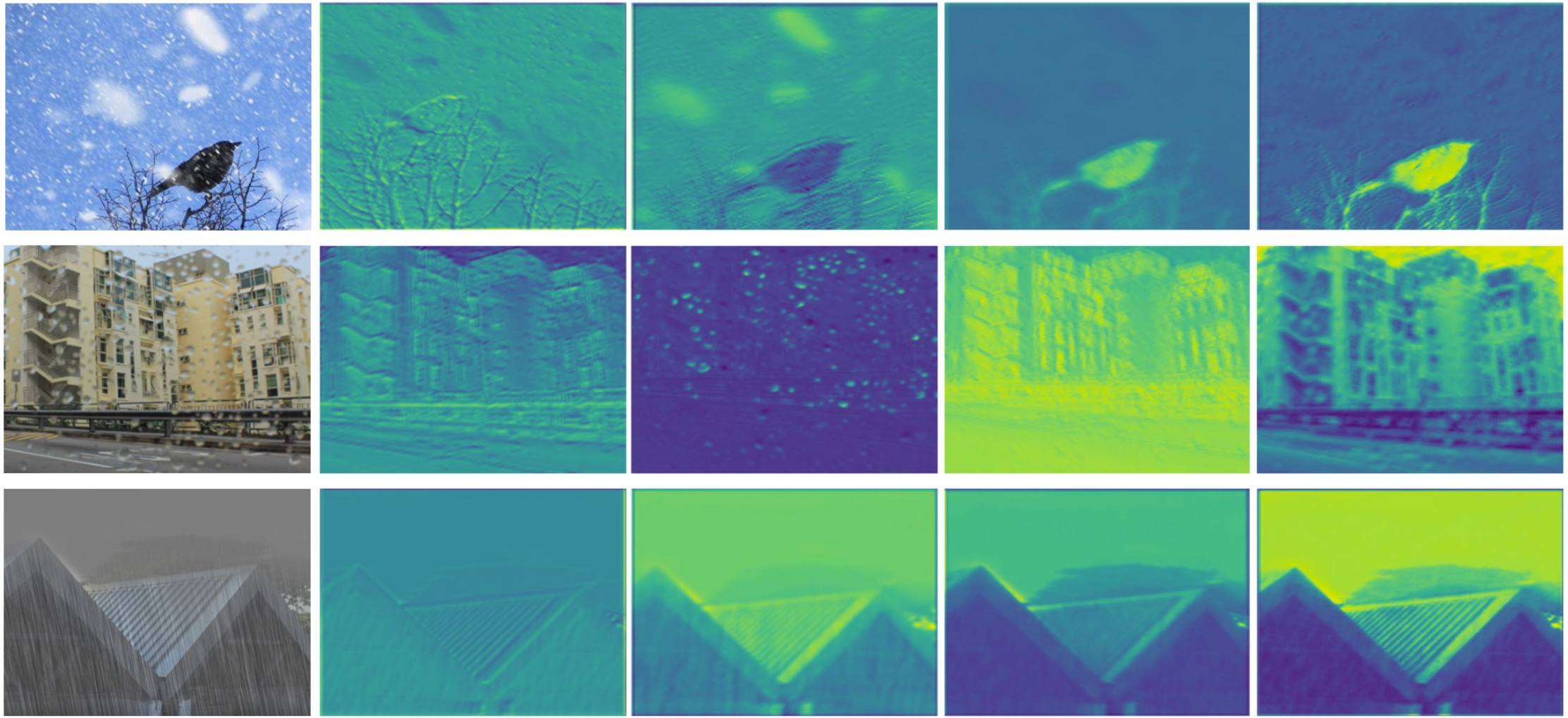
Restoration Feature Aggregation



$$\mathbf{Q} = \mathbf{M}\mathbf{W}^{\mathrm{q}_2}, \quad \mathbf{K} = \hat{\mathbf{X}}^{\mathrm{int}}\mathbf{W}^{\mathrm{k}_2}, \quad \mathbf{V} = \hat{\mathbf{X}}^{\mathrm{int}}\mathbf{W}^{\mathrm{v}_2}, \quad (6)$$

$$\hat{\mathbf{X}} = \mathrm{FFN}^{\mathrm{int}}(\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})), \quad (7)$$

Visualization



(a) $\mathbf{I}^{\mathrm{d}}$ (b) $\mathbf{X}$ (c) $\mathbf{M}$ (d) $\hat{\mathbf{X}}^{\mathrm{int}}$ (e) $\hat{\mathbf{X}}$

Loss function

$$\mathcal{L}_{\text{char}} = \sqrt{\|\mathbf{I}^{\text{c}} - \hat{\mathbf{I}}^{\text{c}}\|^2 + \varepsilon^2}\,, \tag{8}$$

$$\mathcal{L}_{\text{edge}} = \sqrt{\|\nabla\mathbf{I}^{\text{c}} - \nabla\hat{\mathbf{I}}^{\text{c}}\|^2 + \varepsilon^2}\,, \tag{9}$$

- Charbonnier loss
- Gradient-level edge loss

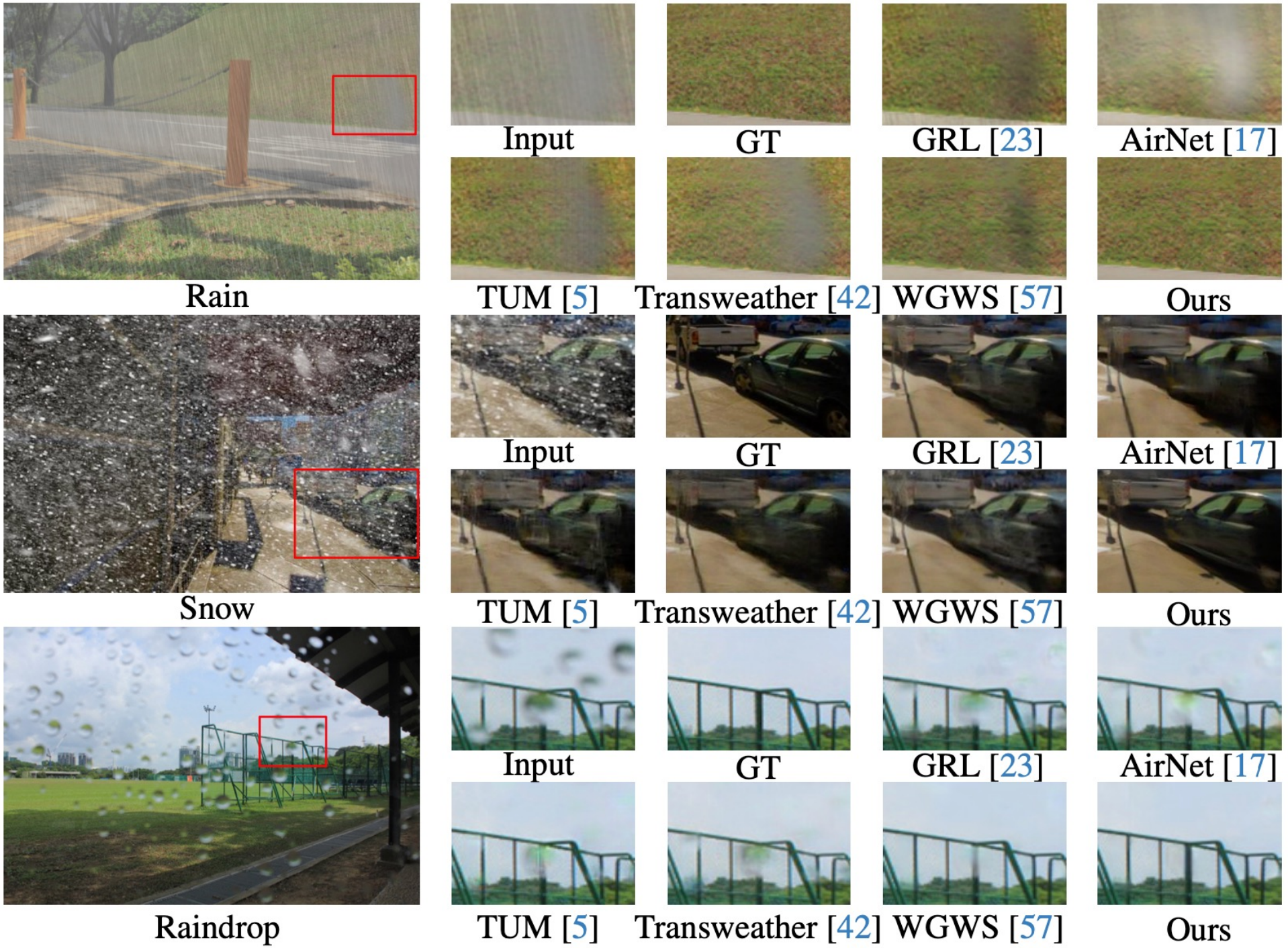
Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Datasets

- Synthetic dataset
  - All-weather dataset
  - 18609 training images and 17609 testing images
  - rain, snow, and raindrops
- Real dataset
  - WeatherStream dataset
  - 176100 training images and 11400 testing images
  - rain, snow, and fog

Result

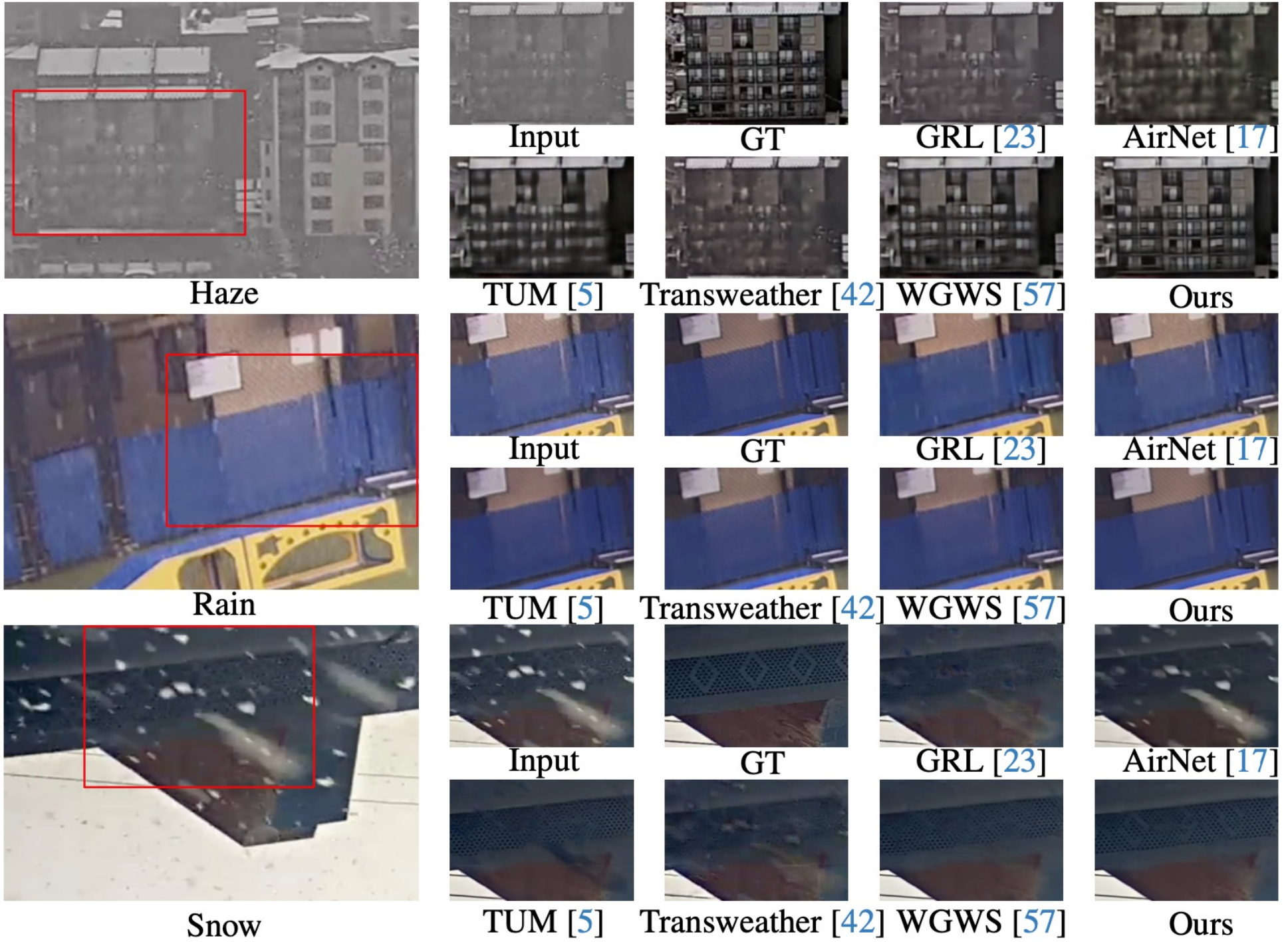

Input
GT
GRL [23]
AirNet [17]
Rain
TUM [5]
Transweather [42]
WGWS [57]
Ours
Input
GT
GRL [23]
AirNet [17]
Snow
TUM [5]
Transweather [42]
WGWS [57]
Ours
Input
GT
GRL [23]
AirNet [17]
Raindrop
TUM [5]
Transweather [42]
WGWS [57]
Ours

Result

Table 1. *Quantitative comparison on the All-weather dataset. We respectively color the best and the second-best methods in red and blue.*

Type	Method	Rain		Snow		Raindrop		Average	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
	BestT + VL	24.33	0.860	28.47	0.872	29.23	0.895	27.34	0.876
	BestT + GT	27.04	0.913	30.61	0.900	31.63	0.936	29.76	0.916
General	MPRNet [51]	23.08	0.839	27.69	0.849	28.75	0.879	26.51	0.856
	NAFNet [3]	23.21	0.840	27.68	0.847	28.90	0.890	26.60	0.859
	Uformer [45]	22.93	0.835	27.50	0.838	28.51	0.871	26.31	0.848
	Restormer [52]	23.37	0.845	27.81	0.850	29.10	0.890	26.76	0.862
	GRL [23]	23.31	0.842	27.79	0.849	29.05	0.888	26.72	0.860
All-in-One	All-in-One [21]	24.71	0.898	28.33	0.882	31.12	0.927	28.05	0.902
	AirNet [17]	23.12	0.837	27.92	0.858	28.23	0.892	26.42	0.862
	TUM [5]	23.92	0.855	29.27	0.884	30.75	0.912	27.98	0.884
	Transweather [42]	23.18	0.841	27.80	0.854	28.98	0.902	26.65	0.866
	WDiff [32]	26.18	0.907	29.69	0.893	29.71	0.911	28.53	0.904
	WGWS [57]	25.31	0.901	29.71	0.894	31.31	0.932	28.78	0.909
	Ours	26.92	0.912	30.79	0.905	31.54	0.933	29.75	0.916

Result


Input
GT
GRL [23]
AirNet [17]
Haze
TUM [5]
Transweather [42]
WGWS [57]
Ours
Input
GT
GRL [23]
AirNet [17]
Rain
TUM [5]
Transweather [42]
WGWS [57]
Ours
Input
GT
GRL [23]
AirNet [17]
Snow
TUM [5]
Transweather [42]
WGWS [57]
Ours

Result

Table 2. *Quantitative comparison on the WeatherStream dataset. We color the best and the second-best methods in red and blue.*

Type	Method	Rain		Haze		Snow		Average	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
	BestT + VL	21.20	0.781	21.60	0.755	20.32	0.772	21.04	0.769
	BestT + GT	23.95	0.810	22.97	0.804	22.70	0.828	23.21	0.814
General	MPRNet [51]	21.50	0.791	21.73	0.763	20.74	0.801	21.32	0.785
	NAFNet [3]	23.01	0.803	22.20	0.803	22.11	0.826	22.44	0.811
	Uformer [45]	22.25	0.791	18.81	0.763	20.94	0.801	20.67	0.785
	Restormer [52]	23.67	0.804	22.90	0.803	22.51	0.828	22.86	0.812
	GRL [23]	23.75	0.805	22.88	0.802	22.59	0.829	23.07	0.812
All-in-One	All-in-One [21]	-	-	-	-	-	-	-	-
	AirNet [17]	22.52	0.797	21.56	0.770	21.44	0.812	21.84	0.793
	TUM [5]	23.22	0.795	22.38	0.805	22.25	0.827	22.62	0.809
	Transweather [42]	22.21	0.772	22.55	0.774	21.79	0.792	22.18	0.779
	WGWS [57]	23.80	0.807	22.78	0.800	22.72	0.831	23.10	0.813
	Ours	24.42	0.818	23.11	0.809	23.12	0.838	23.55	0.822

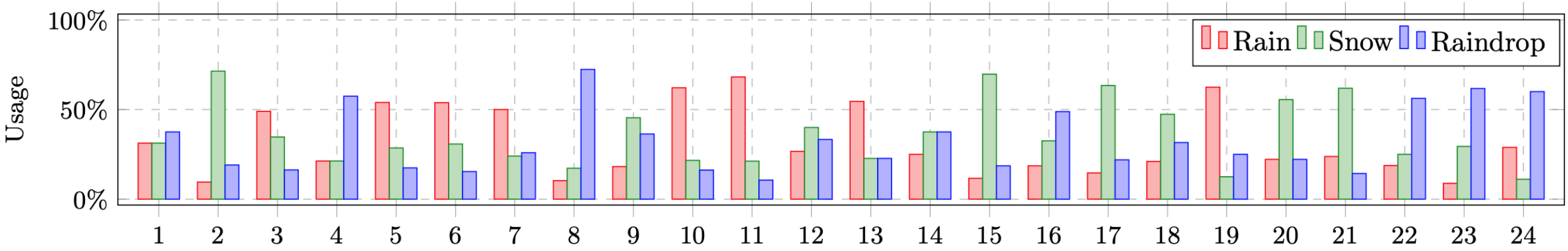
Ablation study

Table 3. *The effectiveness of our model components.*

DMM	TER	RFA	PSNR	SSIM
✗	✓	✗	27.93	0.882
✗	✗	✓	28.37	0.889
✓	✓	✗	28.25	0.890
✓	✗	✓	29.11	0.902
✗	✓	✓	28.55	0.895
✓	✓	✓	29.75	0.916

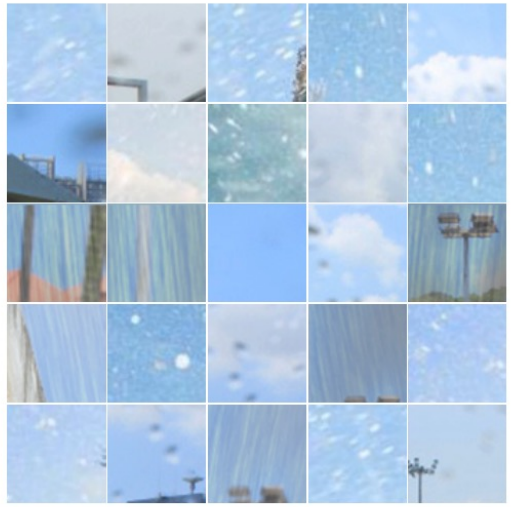
Method	**Slight**		**Moderate**		**Heavy**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
AirNet	27.59	0.895	26.49	0.865	24.50	0.818
WGWS	29.74	0.921	28.77	0.910	27.14	0.886
Ours	30.46	0.925	29.78	0.918	28.38	0.899

Ablation study

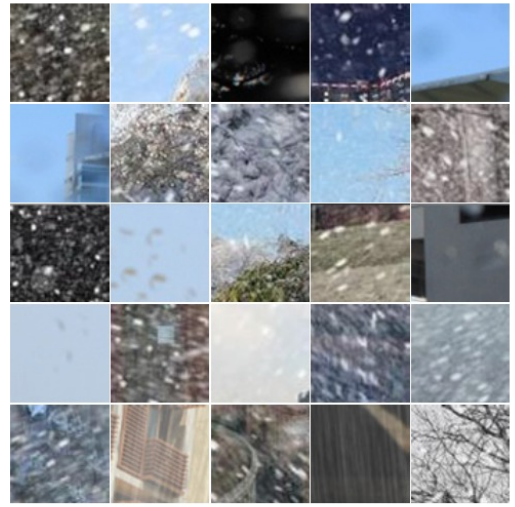


(a) Expert 8

(b) Expert 11



(c) Expert 12



(d) Expert 15

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Conclusion

- Key insight is to leverage a **pre-trained vision-language** model to reason diverse weather-specific knowledge in a degraded image.

- Use this knowledge to restore a clean image with **three modules**: degradation map measurement, Top-K expert restoration, and restoration feature aggregation.

- Experiments on standard benchmark datasets demonstrate that our method **outperforms past works** by a large margin.

HazeCLIP: Towards Language Guided Real-World Image Dehazing

Ruiyi Wang, Wenhao Li, Xiaohong Liu, *Member, IEEE*, Chunyi Li, Zicheng Zhang,
Xiongkuo Min, *Member, IEEE*, and Guangtao Zhai, *Senior Member, IEEE*

arXiv 2024

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Introduction

- Existing methods have achieved remarkable image dehazing performance on synthetic datasets. However, they **struggle with real-world** hazy images due to domain shift.

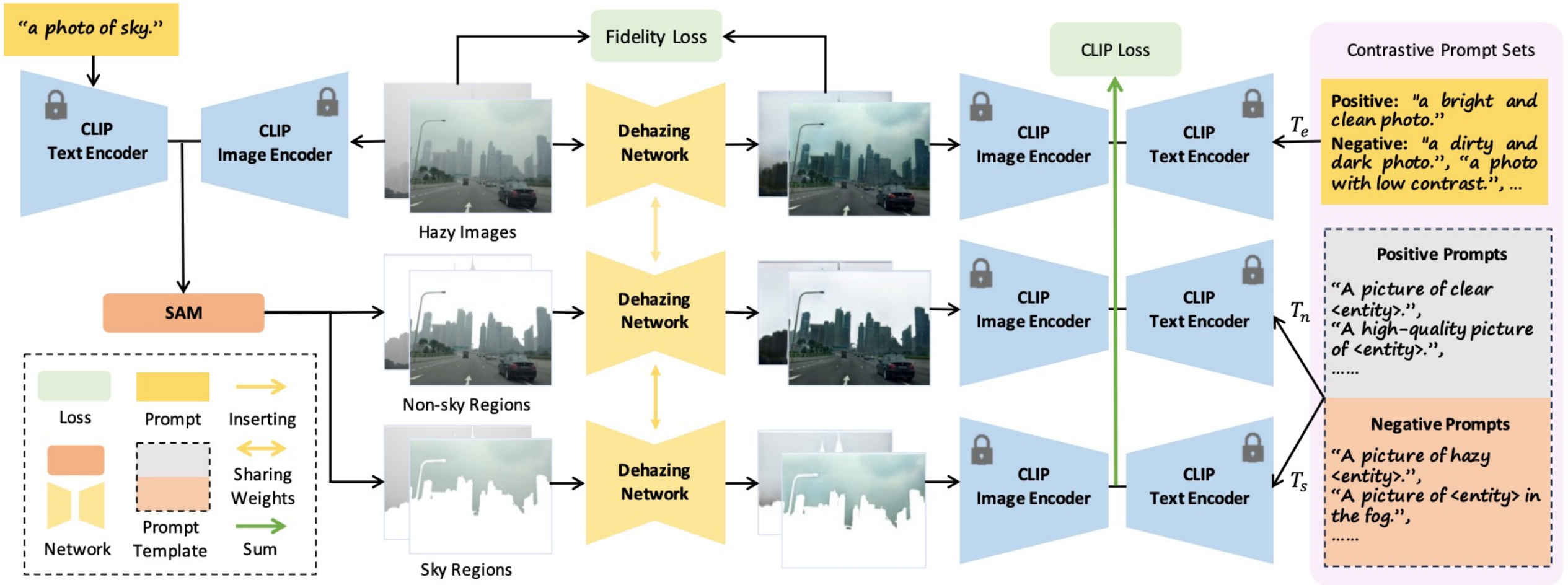
- **Language-guided adaptation** framework designed to enhance the real-world performance of pre-trained dehazing networks.

- Combined with a **region-specific** dehazing technique and tailored **prompt sets**, CLIP model accurately identifies hazy areas, providing a **high-quality prior** that guides the fine-tuning process of pre-trained networks.

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

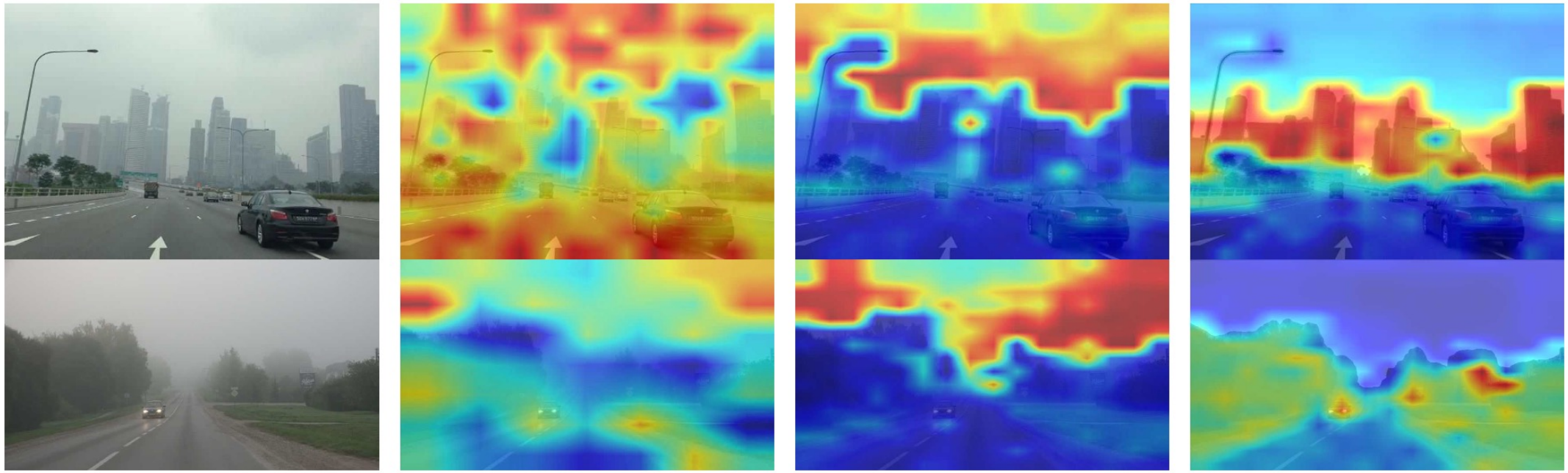
Framework


"a photo of sky."
CLIP Text Encoder
CLIP Image Encoder
Fidelity Loss
Hazy Images
Dehazing Network
CLIP Loss
CLIP Image Encoder
CLIP Text Encoder
T_e
Contrastive Prompt Sets
Positive: "a bright and clean photo."
Negative: "a dirty and dark photo.", "a photo with low contrast.", ...
SAM
Non-sky Regions
Dehazing Network
CLIP Image Encoder
CLIP Text Encoder
T_n
Positive Prompts
"A picture of clear <entity>.",
"A high-quality picture of <entity>.",
.....
Sky Regions
Dehazing Network
CLIP Image Encoder
CLIP Text Encoder
T_s
Negative Prompts
"A picture of hazy <entity>.",
"A picture of <entity> in the fog.",
.....
Loss
Prompt
Inserting
Sharing Weights
Network
Prompt Template
Sum

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Region-Specific Dehazing



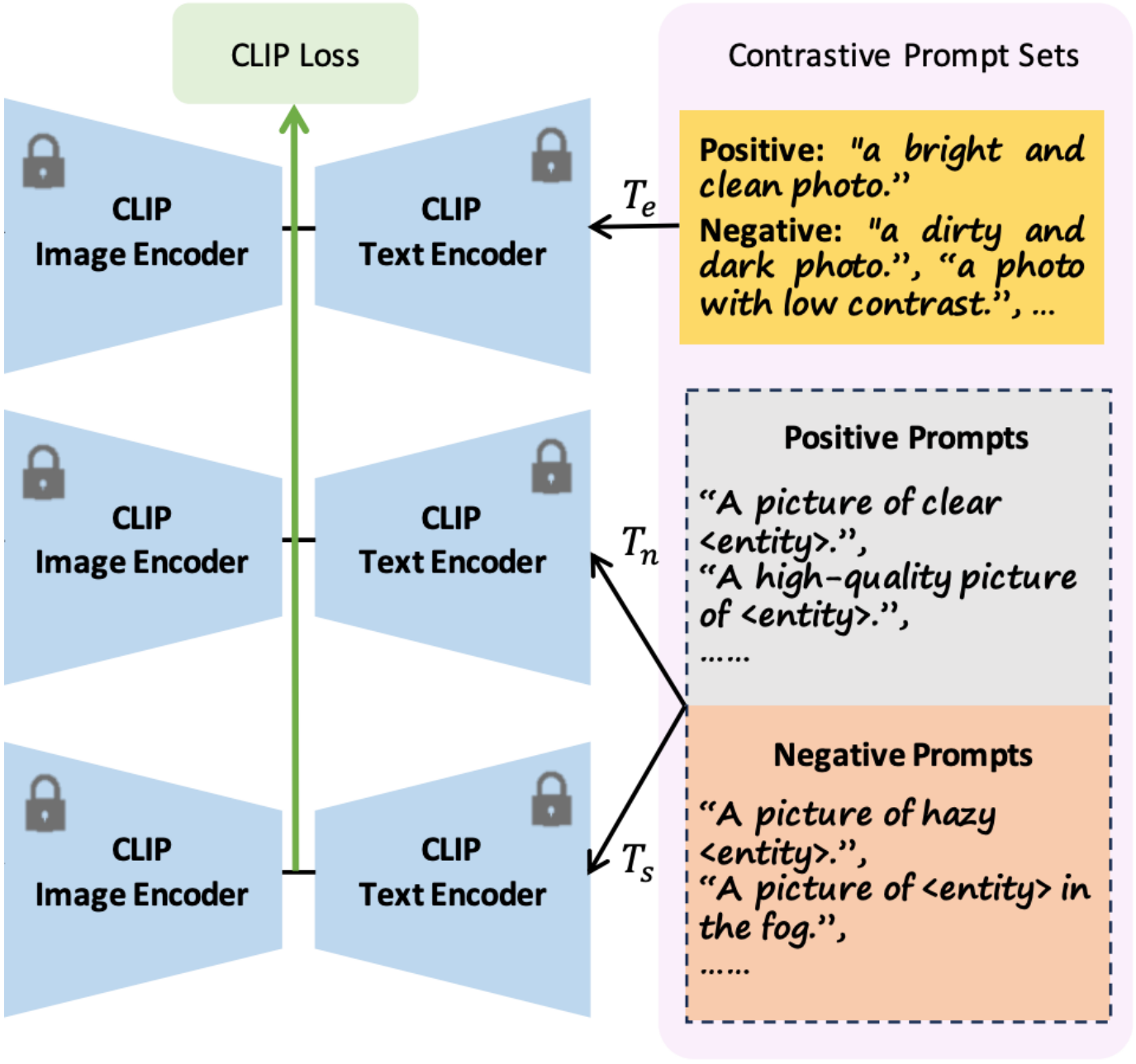
(a) (b) (c) (d)

- **Training data bias** where images with haze captions often feature grey skies, yields excessively high haze similarity scores for sky areas
- Propose a **region-specific** dehazing technique that separately processes the sky and non-sky regions.

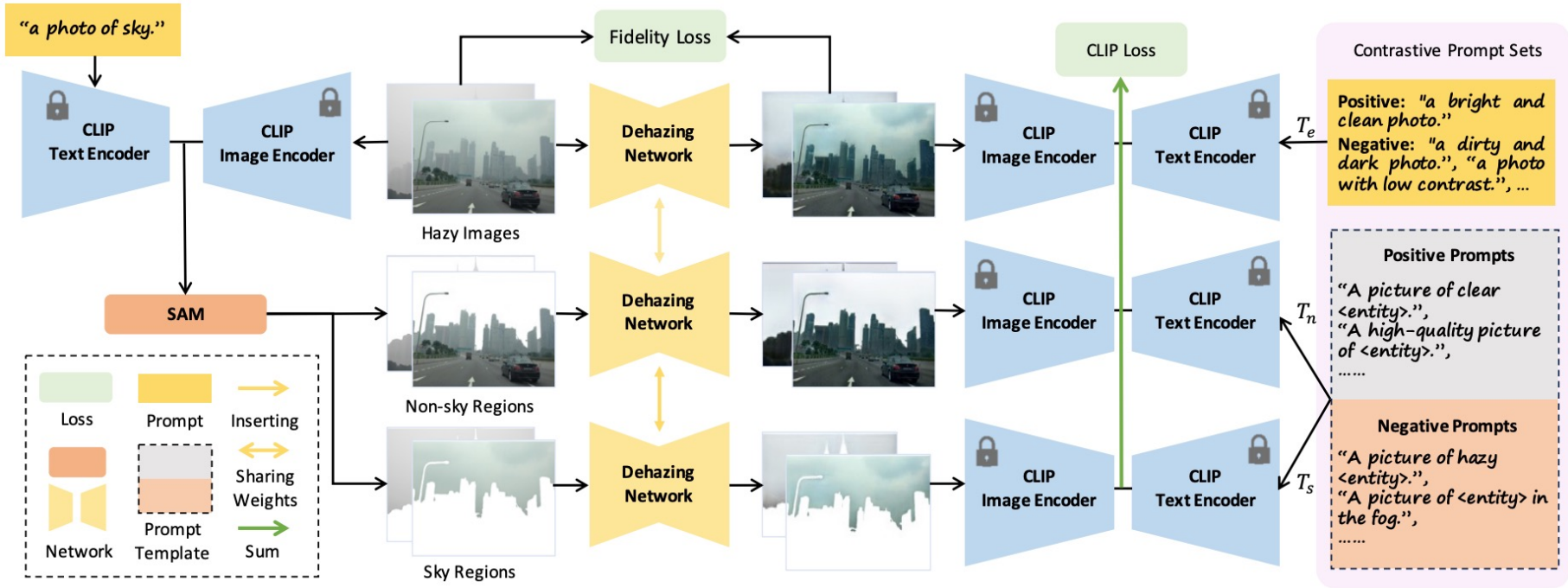
Contrastive Prompt

$$\mathcal{L}_T(I) = \frac{e^{cos(\Phi_i(I),\Phi_t(T^p))}}{\sum_{j\in\{p,n_1,\ldots,n_K\}} e^{cos(\Phi_i(I),\Phi_t(T^j))}}$$

$$T \in \{T_s, T_n, T_e\}$$



Synthetic-to-Real Adaptation



- CLIP guidance loss

$$\mathcal{L}_c(I) = \mathcal{L}_{T_s}(\mathcal{M}(I_s)) + \mathcal{L}_{T_n}(\mathcal{M}(I_n)) + \lambda_1 \cdot \mathcal{L}_{T_e}(\mathcal{M}(I)), \quad (2)$$

- Fidelity loss

$$\mathcal{L}_f(I) = \Sigma_{l=0}^{4} \alpha_l \cdot \|\Phi_i^l(\mathcal{M}(I)) - \Phi_i^l(I)\|_2. \quad (3)$$

- Total loss

$$\mathcal{L}(I) = \mathcal{L}_c(I) + \lambda_2 \cdot \mathcal{L}_f(I).$$

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Experiment setup

- Pre-training
  - **Synthetic** data
  - RIDCP
- Fine-tuning
  - **Real-world** data
  - URHI split from the RESIDE dataset
- Testing
  - **Real-world** data
  - RTTS split from the RESIDE dataset

Result

Result

Method	FADE↓	BRISQUE↓	NIMA↑	MOS↑
Hazy image	2.484	37.011	4.3250	-
Dehamer [40]	1.895	33.866	3.8663	2.78
DAD [27]	1.130	32.727	4.0055	3.13
PSD [30]	**0.920**	25.239	4.3459	3.20
D4 [41]	1.358	33.206	3.7239	2.48
RIDCP [31]	0.944	**18.782**	4.4267	**3.57**
DCMPNet [42]	1.921	32.520	**4.4351**	2.85
HazeCLIP	**0.638**	**18.567**	**4.5510**	**3.60**

Ablation study

Metric	setting (a)	setting (b)	setting (c)	full version
FADE ↓	1.091	0.857	0.695	**0.638**
BRISQUE ↓	27.309	20.499	22.376	**18.567**
NIMA ↑	4.3961	4.4587	4.3965	**4.5510**

- (a) without HazeCLIP adaptation
- (b) without the region-specific dehazing technique
  - one dehazing prompt set applies to the whole image
- (c) without the enhancing prompt set

Ablation study

Method	FADE↓	BRISQUE↓	NIMA↑
GDN [24]	1.470	29.448	4.2502
GDN+HazeCLIP	**0.976**	**21.352**	**4.3252**
FFANet [16]	1.153	26.233	4.2817
FFANet+HazeCLIP	**0.913**	**19.522**	**4.4019**
MSBDN [17]	1.091	27.309	4.3961
MSBDN+HazeCLIP	**0.638**	**18.567**	**4.5510**

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Conclusion

- Generalize dehazing networks pre-trained on **synthetic data to real-world** applications.

- Employing the **region-specific** dehazing technique and specially **designed prompt sets**, HazeCLIP accurately detect hazy regions and guide the fine-tuning process with precision.

- **Versatile fine-tuning framework**, compatible with various dehazing networks, highlighting its practical value.