Controlling Vision-Language Models for Multi-Task Image Restoration

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, Thomas B. Schön

Uppsala University

Project Page:
https://algolzw.github.io/daclip-uir/index.html

Multi-Task Image Restoration

Key Idea: Perform restoration across multiple tasks and datasets with a single model.

- Better generalization for different degradation types.
- Efficient deployment of real-world applications.

Multi-Task Image Restoration

UPPSALA UNIVERSITET

Key Idea: Perform restoration across multiple tasks and datasets with a single model.

- Better generalization for different degradation types.
- Efficient deployment of real-world applications.

Multi-Task Image Restoration

UPPSALA
UNIVERSITET

Key Idea: Perform restoration across multiple tasks and datasets with a single model.

- Better generalization for different degradation types.
- Efficient deployment of real-world applications.

Examples of Multi-Task Image Restoration

Image dehazing

Image deblurring

Face inpainting

JPEG artifact deduction

A Single Restoration Model for
All Degradations

Image desnowing

Low-light image enhancement

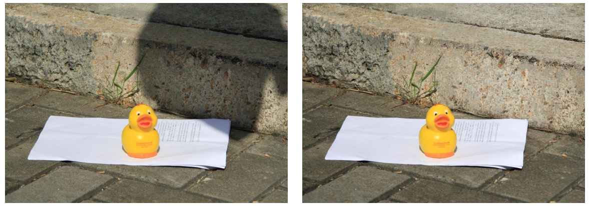


Image shadow removal

Image denoising

Image raindrop removal

Image deraining

Solutions and Challenges

Common solutions:

- Train individual modules for degradation classification
    - AirNet[1] trains an extra encoder to differentiate degradation types using contrastive learning.
    - PromptIR[2] employs a visual prompt module to guide the restoration for different tasks.
    - ***They are limited to a small number of degradation types and the specific datasets!***

Challenges:

- *Inaccurate prediction due to unbalanced datasets of different tasks.*
- *No proper way to utilize the degradation for image restoration.*

[1] Li, Boyun, et al. "*All-in-one image restoration for unknown corruption*". CVPR 2022.
[2] Potlapalli, Vaishnav, et al. "*Promptir: Prompting for all-in-one blind image restoration*". NeurIPS 2023.

Solutions and Challenges

UPPSALA
UNIVERSITET

Common solutions:

- Train individual modules for degradation classification
  - AirNet[1] trains an extra encoder to differentiate degradation types using contrastive learning.
  - PromptIR[2] employs a visual prompt module to guide the restoration for different tasks.
  - ***They are limited to a small number of degradation types and the specific datasets!**

Challenges:

- *Inaccurate prediction due to unbalanced datasets of different tasks.*
- *No proper way to utilize the degradation for image restoration.*

Our Idea: ***Large-scale pretrained vision-language models!!***

[1] Li, Boyun, et al. "*All-in-one image restoration for unknown corruption*". CVPR 2022.
[2] Potlapalli, Vaishnav, et al. "*Promptir: Prompting for all-in-one blind image restoration*". NeurIPS 2023.

Solutions and Challenges

UPPSALA UNIVERSITET

Common solutions:

- Train individual modules for degradation classification
  - AirNet[1] trains an extra encoder to differentiate degradation types using contrastive learning.
  - PromptIR[2] employs a visual prompt module to guide the restoration for different tasks.
  - ****They are limited to a small number of degradation types and the specific datasets!***

Challenges:

- *Inaccurate prediction due to unbalanced datasets of different tasks.*
- *No proper way to utilize the degradation for image restoration.*

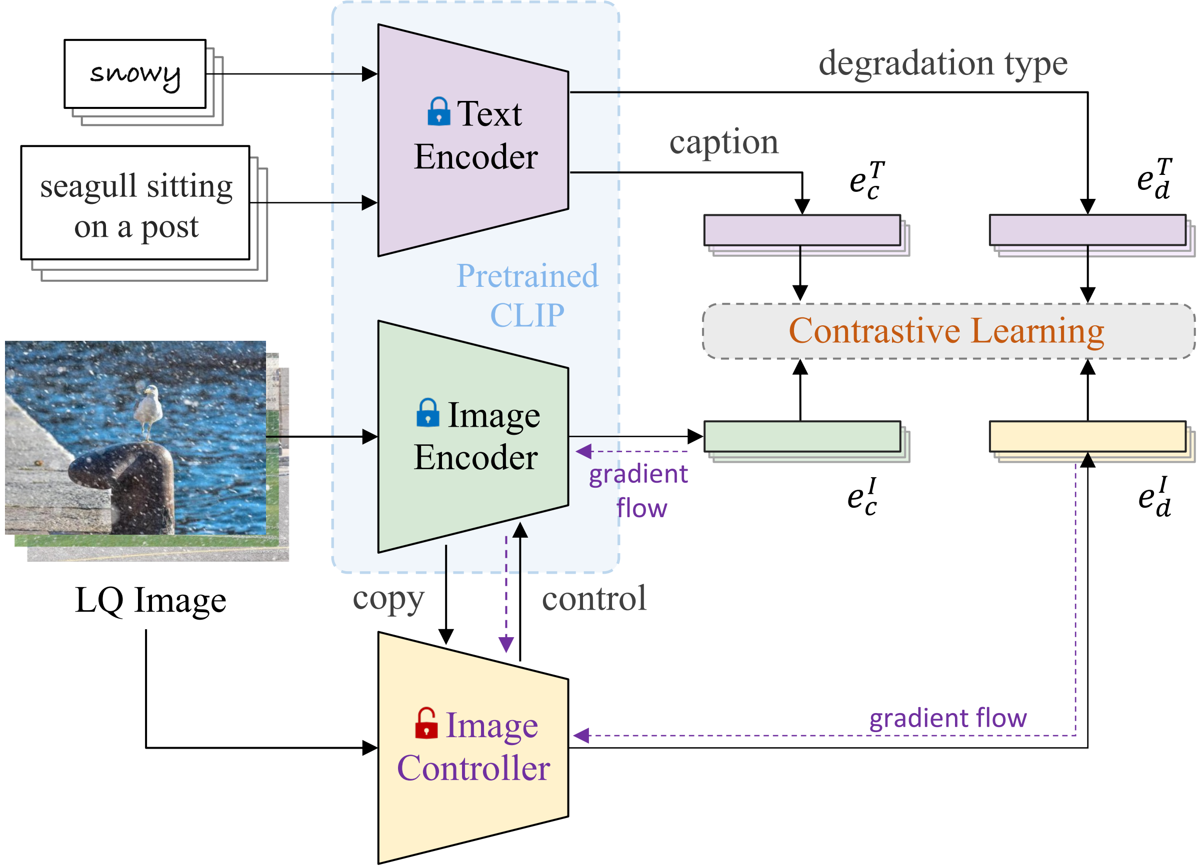
Our Idea: ***Large-scale pretrained vision-language models!!***

How?

[1] Li, Boyun, et al. "*All-in-one image restoration for unknown corruption*". CVPR 2022.
[2] Potlapalli, Vaishnav, et al. "*Promptir: Prompting for all-in-one blind image restoration*". NeurIPS 2023.

Degradation-aware CLIP (DA-CLIP)

- Controlling a pretrained vision-language model, i.e. CLIP[3], for:
  1. <u>*Degradation embedding*</u> prediction.
  2. <u>*Content embedding*</u> prediction.

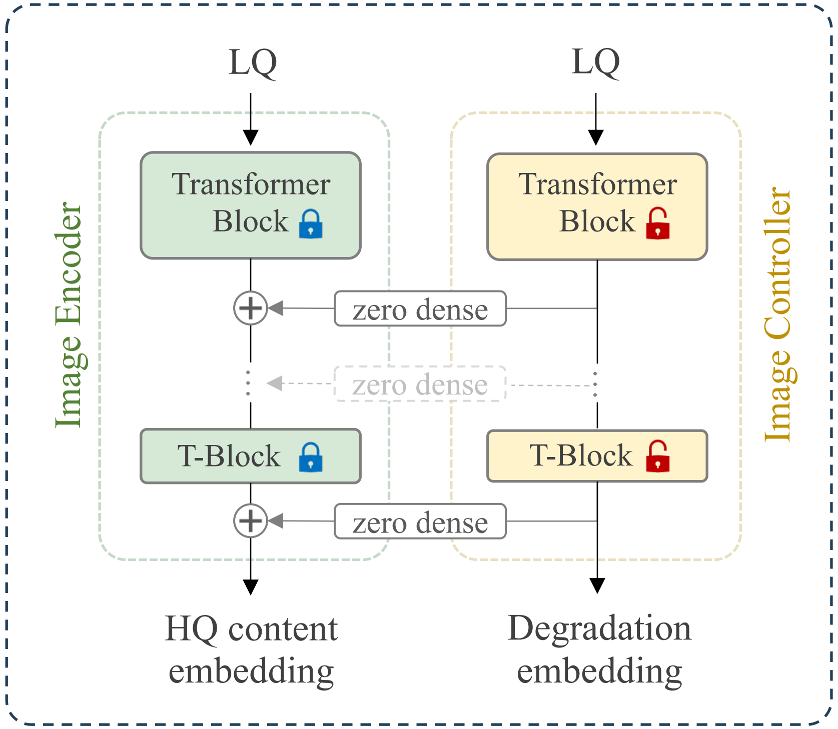


[3] Radford, Alec, et al. "*Learning transferable visual models from natural language supervision.*" ICML 2021.

Degradation-aware CLIP (DA-CLIP)

UPPSALA UNIVERSITET

- Controlling a pretrained vision-language model, i.e. CLIP[3], for:
  1. *Degradation embedding* prediction.
  2. *Content embedding* prediction.



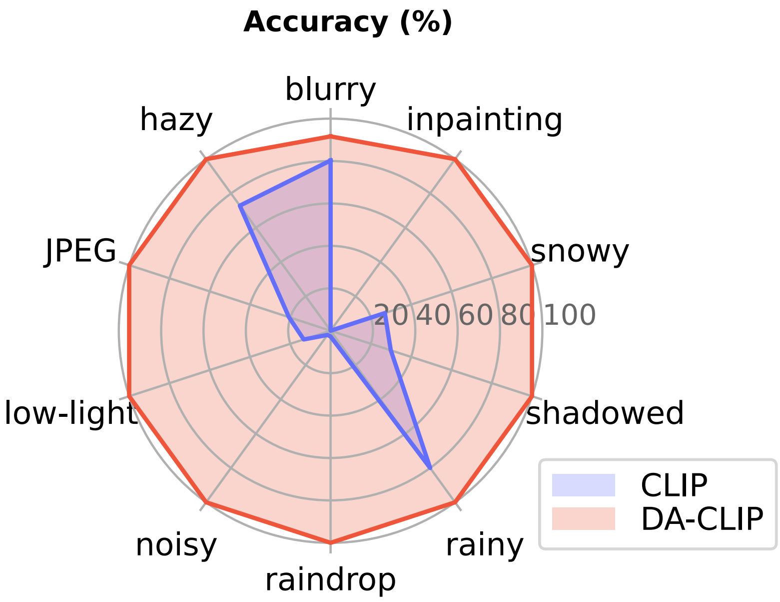
Example: Controller for ViT-based image encoder

[3] Radford, Alec, et al. "*Learning transferable visual models from natural language supervision.*" ICML 2021.

Degradation-aware CLIP (DA-CLIP)

UPPSALA UNIVERSITET

- Controlling a pretrained vision-language model, i.e. CLIP[3], for:
  1. *Degradation embedding* prediction.
  2. *Content embedding* prediction.



CLIP vs. DA-CLIP on 10 Degradation Types

[3] Radford, Alec, et al. "*Learning transferable visual models from natural language supervision.*" ICML 2021.

Image Restoration with DA-CLIP

UPPSALA UNIVERSITET

- Integrating *content embeddings* into U-Net with **cross-attention**.
- Integrating *degradation embeddings* into U-Net with **visual prompt**.

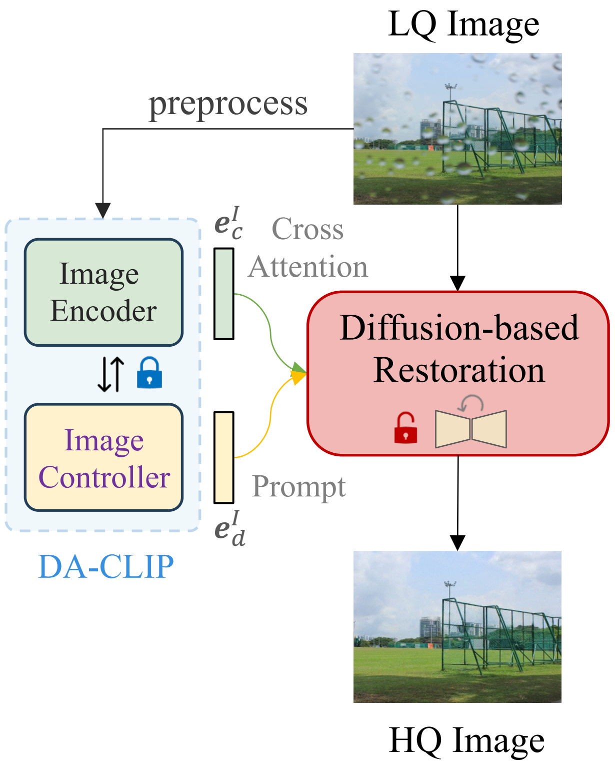
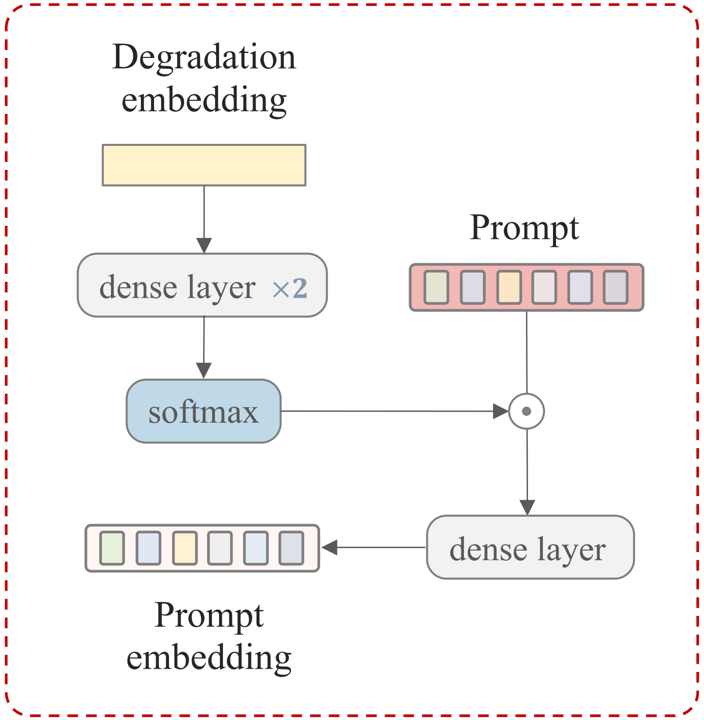


Image Restoration with DA-CLIP

UPPSALA UNIVERSITET

- Integrating *content embeddings* into U-Net with **cross-attention**
- Integrating *degradation embeddings* into U-Net with **visual prompt**



Prompt for degradation embeddings

Dataset Construction

UPPSALA
UNIVERSITET

HQ Image

LQ Image

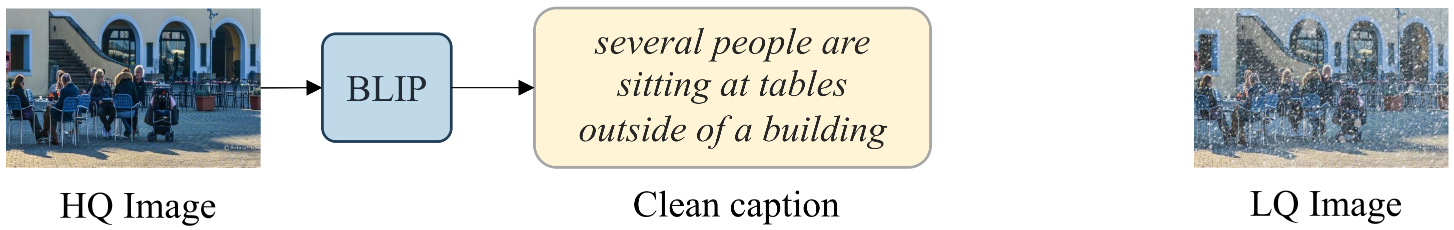
Dataset Construction

HQ Image

BLIP

LQ Image

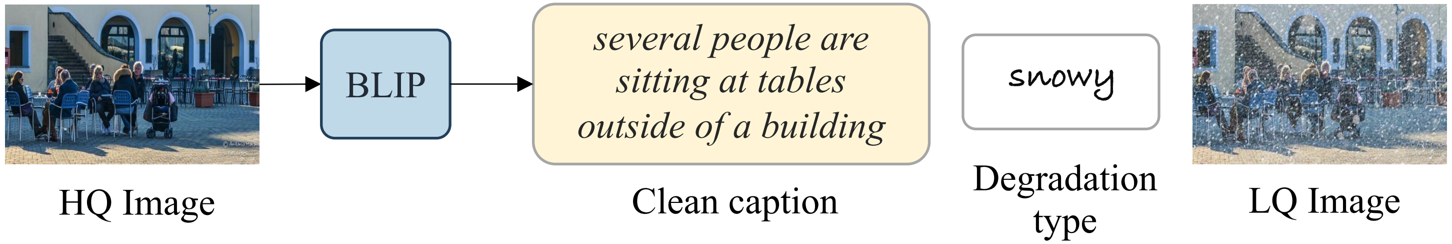
Dataset Construction



UPPSALA UNIVERSITET

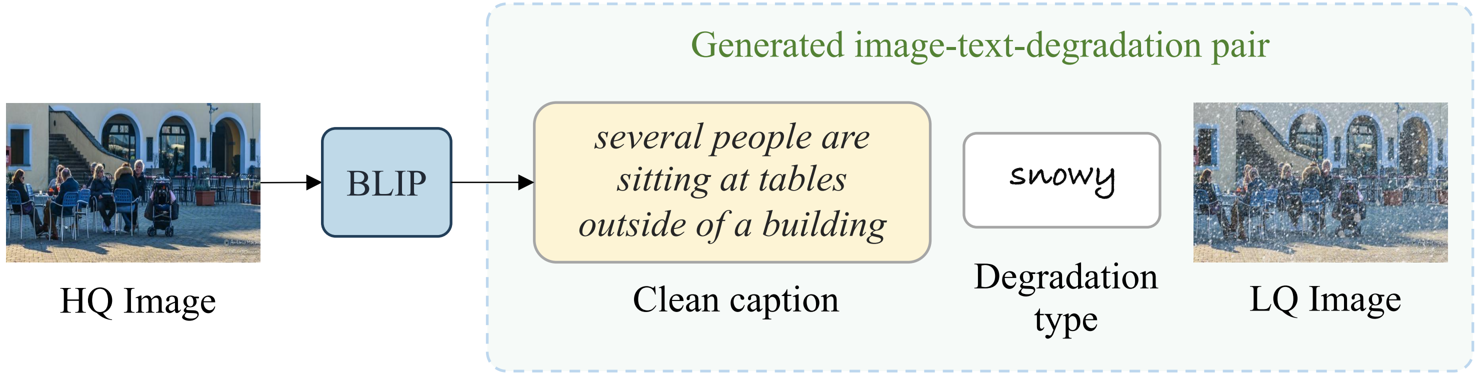
Dataset Construction

UPPSALA UNIVERSITET



Dataset Construction

UPPSALA UNIVERSITET



Dataset Construction

UPPSALA UNIVERSITET

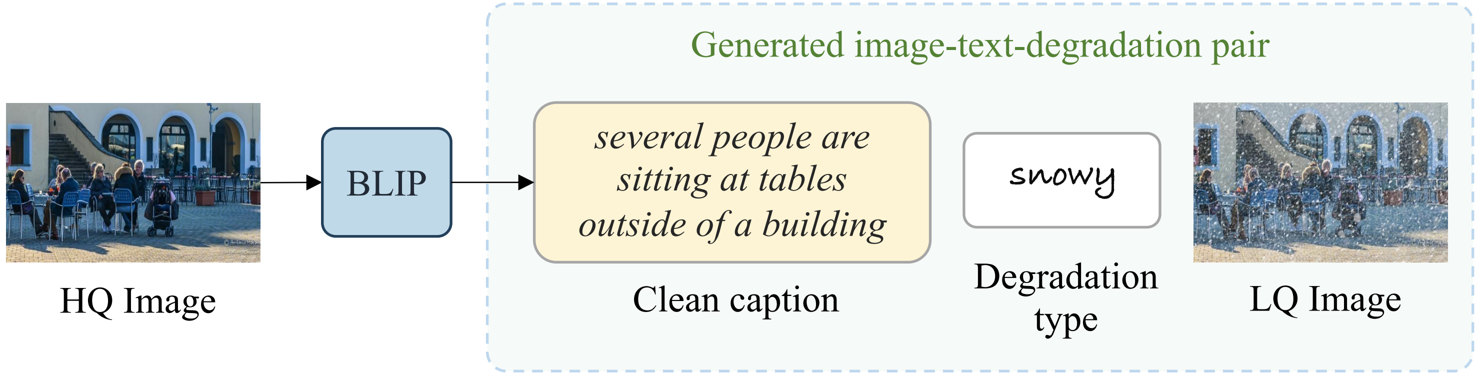


Table 1: Details of the collected training and testing datasets with different image degradation types.

Dataset	Blurry	Hazy	JPEG	Low-light	Noisy	Raindrop	Rainy	Shadowed	Snowy	Inpainting
#Train	2 103	6 000	3 550	485	3 550	861	1 800	2 680	1 872	29 900
#Test	1 111	1 000	29	15	68	58	100	408	601	100

Experiments

Degradation-Specific Image Restoration

a) Deraining
b) Low-light enhancement
c) Deblurring
d) Dehazing

Table 2: Quantitative comparison between our method with other state-of-the-art approaches on four different *degradation-specific* datasets. The best results are marked in boldface.

Method	Distortion		Perceptual	
	PSNR↑	SSIM↑	LPIPS↓	FID↓
JORDER	26.25	0.835	0.197	94.58
PReNet	29.46	0.899	0.128	52.67
MPRNet	30.41	0.891	0.158	61.59
MAXIM	30.81	0.903	0.133	58.72
IR-SDE	31.65	0.904	0.047	18.64
Ours	**33.91**	**0.926**	**0.031**	**11.79**

(a) Deraining results on the Rain100H dataset.

Method	Distortion		Perceptual	
	PSNR↑	SSIM↑	LPIPS↓	FID↓
EnlightenGAN	17.61	0.653	0.372	94.71
MIRNet	**24.14**	0.830	0.250	69.18
URetinex-Net	19.84	0.824	0.237	52.38
MAXIM	23.43	**0.863**	0.098	48.59
IR-SDE	20.45	0.787	0.129	47.28
Ours	23.77	0.830	**0.083**	**34.03**

(b) Low-light enhancement on the LOL dataset.

Method	Distortion		Perceptual	
	PSNR↑	SSIM↑	LPIPS↓	FID↓
DeepDeblur	29.08	0.913	0.135	15.14
DeblurGAN	28.70	0.858	0.178	27.02
DeblurGANv2	29.55	0.934	0.117	13.40
MAXIM	**32.86**	**0.940**	0.089	11.57
IR-SDE	30.70	0.901	0.064	6.32
Ours	30.88	0.903	**0.058**	**6.15**

(c) Deblurring results on the GoPro dataset.

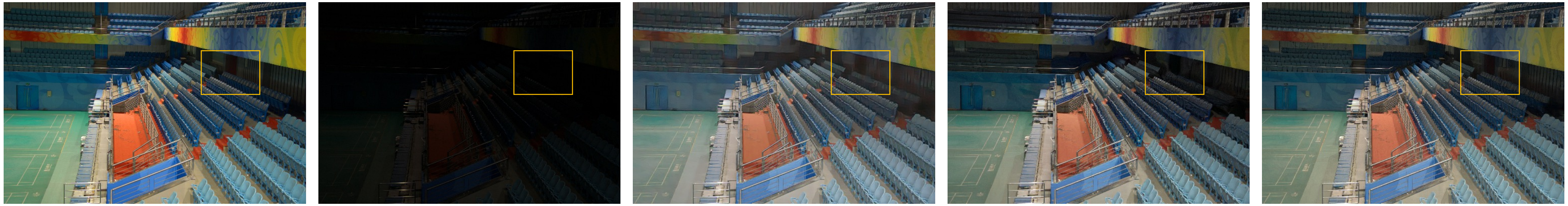
Method	Distortion		Perceptual	
	PSNR↑	SSIM↑	LPIPS↓	FID↓
GCANet	26.59	0.935	0.052	11.52
GridDehazeNet	25.86	0.944	0.048	10.62
DehazeFormer	**30.29**	**0.964**	0.045	7.58
MAXIM	29.12	0.932	0.043	8.12
IR-SDE	25.25	0.906	0.060	8.33
Ours	30.16	0.936	**0.030**	**5.52**

(d) Dehazing results on the RESIDE-6k dataset.

Degradation-Specific Image Restoration

a) Deraining b) Low-light enhancement c) Dehazing d) Deblurring

HQ Image | Rainy Image | MPRNet | IR-SDE | Ours



HQ Image | Low-light Image | URetinexNet | IR-SDE | Ours

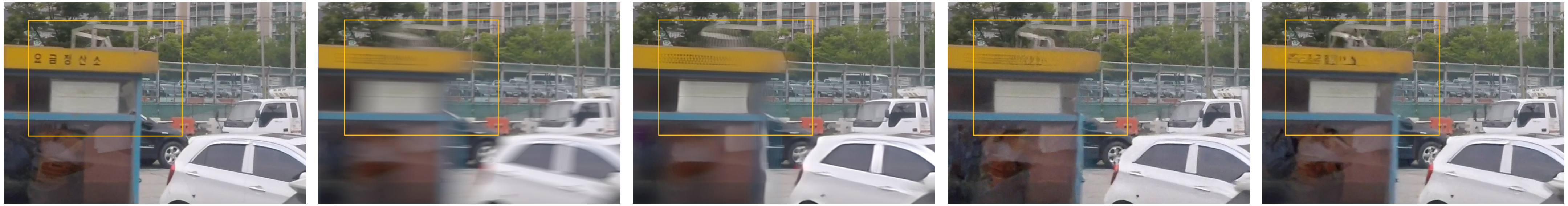
Comparison of our method with other approaches on different degradation-specific tasks.

Degradation-Specific Image Restoration

UPPSALA UNIVERSITET

a) Deraining b) Low-light enhancement c) Dehazing d) Deblurring

HQ Image | Rainy Image | DehazeFormer | IR-SDE | Ours



HQ Image | Low-light Image | DeblurGAN-v2 | IR-SDE | Ours

Comparison of our method with other approaches on different degradation-specific tasks.

Unified Image Restoration

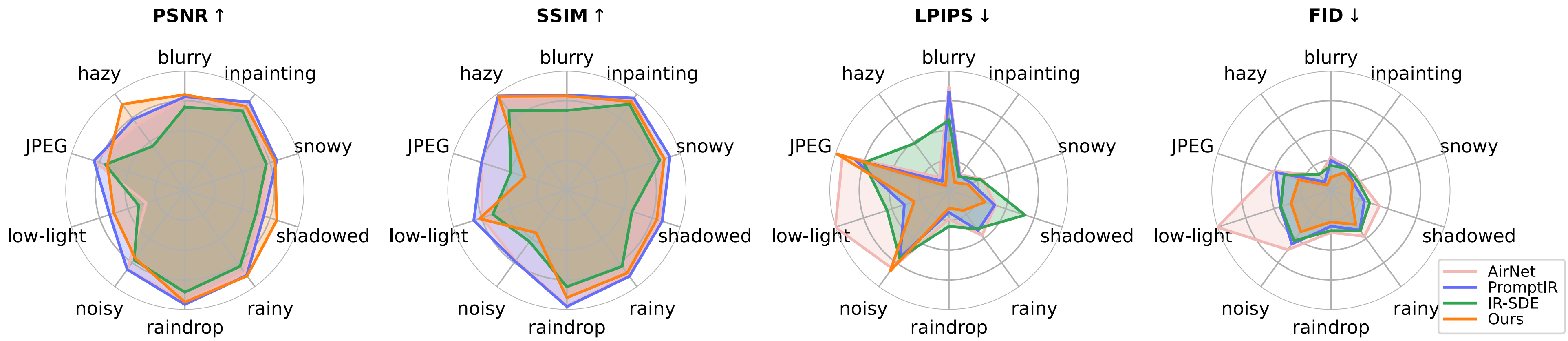


Table 3: Comparison of the average results over ten different datasets on the *unified* image restoration task.

Method	Distortion		Perceptual	
	PSNR↑	SSIM↑	LPIPS↓	FID↓
NAFNet	26.34	0.847	0.159	55.68
NAFNet + Degradation	27.02	0.856	0.146	48.27
NAFNet + DA-CLIP	**27.22**	**0.861**	0.145	47.94
Restormer	26.43	0.850	0.157	54.03
AirNet	25.62	0.844	0.182	64.86
PromptIR	27.14	0.859	0.147	48.26
IR-SDE	23.64	0.754	0.167	49.18
Ours	27.01	0.794	**0.127**	**34.89**

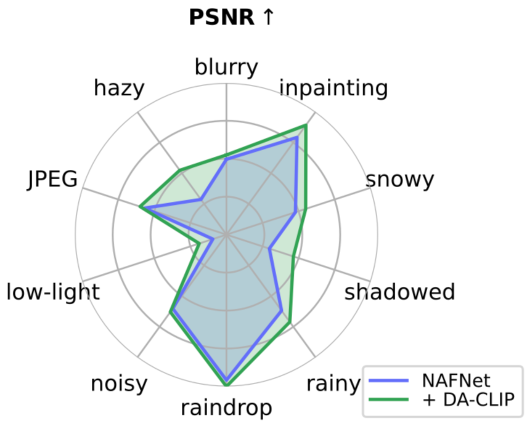
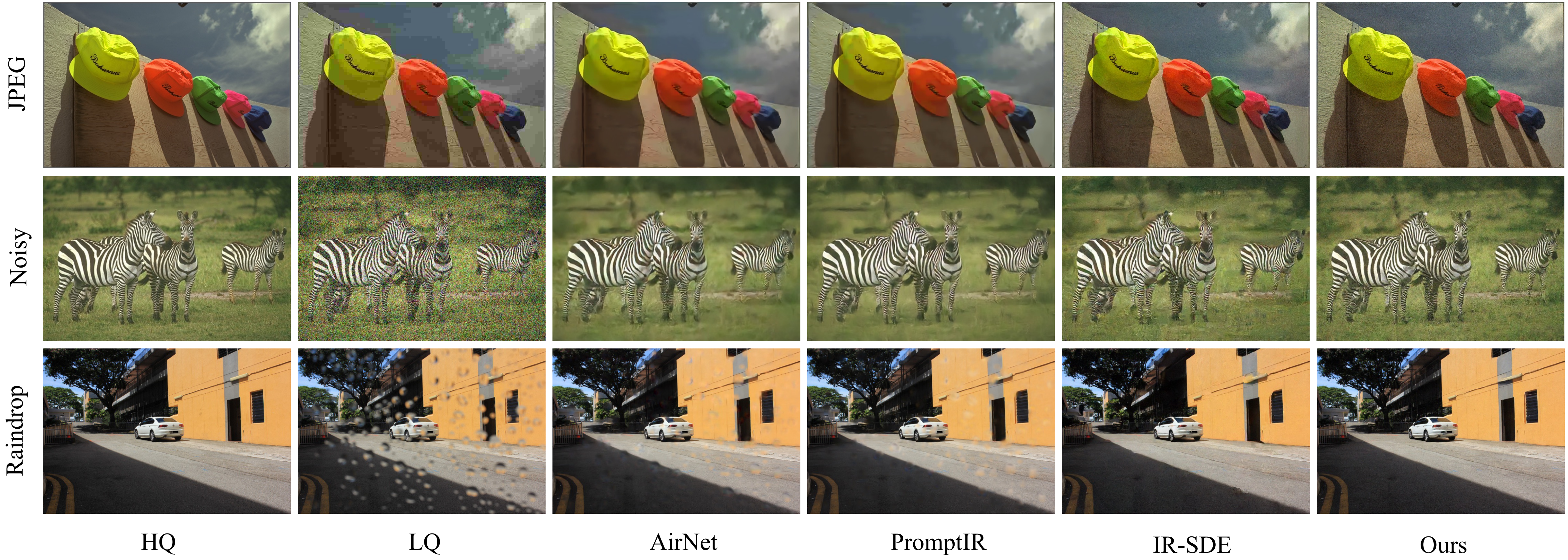


Figure 7: NAFNet with DA-CLIP for *unified* image restoration.

Unified Image Restoration

UPPSALA
UNIVERSITET

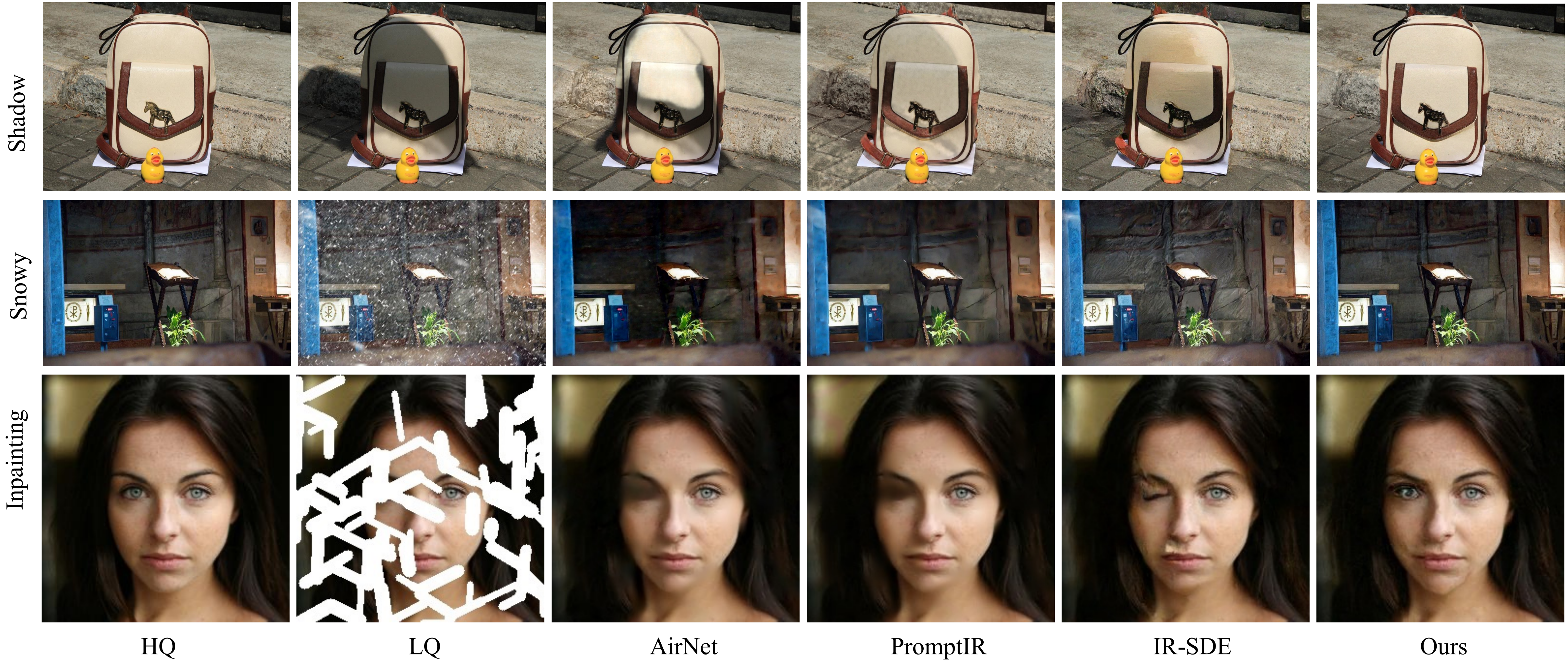
Comparison of our method with other approaches on the unified image restoration.



Unified Image Restoration

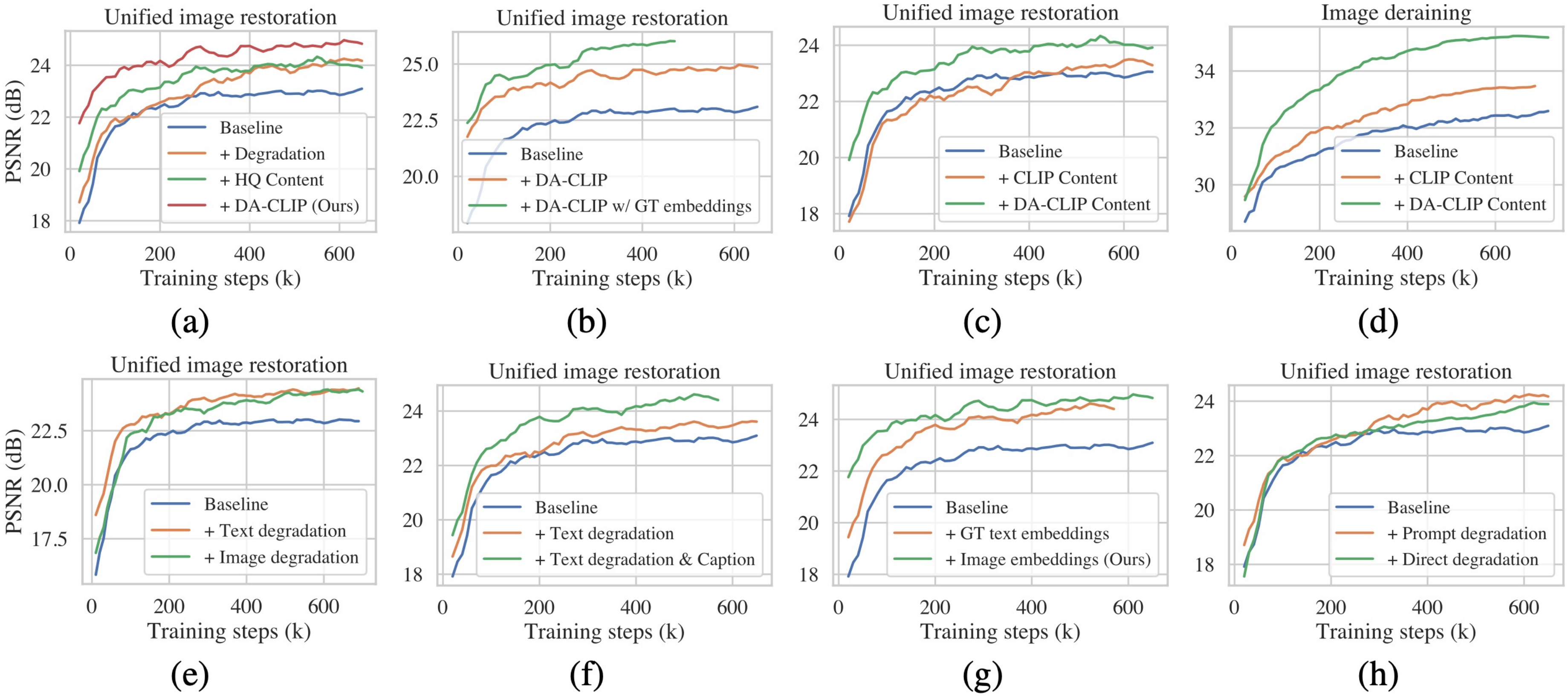
UPPSALA UNIVERSITET

Comparison of our method with other approaches on the unified image restoration.



Discussion & Analysis

UPPSALA
UNIVERSITET



Summary

UPPSALA UNIVERSITET

- We present **DA-CLIP** to leverage pretrained vision-language models for image restoration.
- We use **cross-attention** and **visual prompt** to integrate embeddings into restoration networks.
- We construct **a mixed degradation dataset** with synthetic captions for DA-CLIP training.
- Our approach achieves highly competitive performance on diverse image restoration tasks.

Summary

UPPSALA UNIVERSITET

- We present **DA-CLIP** to leverage pretrained vision-language models for image restoration.
- We use **cross-attention** and **visual prompt** to integrate embeddings into restoration networks.
- We construct **a mixed degradation dataset** with synthetic captions for DA-CLIP training.
- Our approach achieves highly competitive performance on diverse image restoration tasks.

Summary

UPPSALA UNIVERSITET

- We present **DA-CLIP** to leverage pretrained vision-language models for image restoration.
- We use **cross-attention** and **visual prompt** to integrate embeddings into restoration networks.
- We construct **a mixed degradation dataset** with synthetic captions for DA-CLIP training.
- Our approach achieves highly competitive performance on diverse image restoration tasks.

Summary

UPPSALA
UNIVERSITET

- We present **DA-CLIP** to leverage pretrained vision-language models for image restoration.
- We use **cross-attention** and **visual prompt** to integrate embeddings into restoration networks.
- We construct **a mixed degradation dataset** with synthetic captions for DA-CLIP training.
- Our approach achieves highly competitive performance on diverse image restoration tasks.

UPPSALA
UNIVERSITET

Thanks

Project Page:
https://algolzw.github.io/daclip-uir/index.html

Follow-up Paper:
Photo-Realistic Image Restoration in the Wild with Controlled Vision-Language Models