CONTROLLING VISION-LANGUAGE MODELS FOR UNIVERSAL IMAGE RESTORATION

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, Thomas B. Schön
Department of Information Technology, Uppsala University
{ziwei.luo,fredrik.gustafsson,zheng.zhao}@it.uu.se
{jens.sjolund,thomas.schon}@it.uu.se

ABSTRACT

Vision-language models such as CLIP have shown great impact on diverse downstream tasks for zero-shot or label-free predictions. However, when it comes to low-level vision such as image restoration their performance deteriorates dramatically due to corrupted inputs. In this paper, we present a degradation-aware vision-language model (DA-CLIP) to better transfer pretrained vision-language models to low-level vision tasks as a universal framework for image restoration. More specifically, DA-CLIP trains an additional controller that adapts the fixed CLIP image encoder to predict high-quality feature embeddings. By integrating the embedding into an image restoration network via cross-attention, we are able to pilot the model to learn a high-fidelity image reconstruction. The controller itself will also output a degradation feature that matches the real corruptions of the input, yielding a natural classifier for different degradation types. In addition, we construct a mixed degradation dataset with synthetic captions for DA-CLIP training. Our approach advances state-of-the-art performance on both *degradation-specific* and *unified* image restoration tasks, showing a promising direction of prompting image restoration with large-scale pretrained vision-language models. Our code is available at https://github.com/Algolzw/daclip-uir.

1 INTRODUCTION

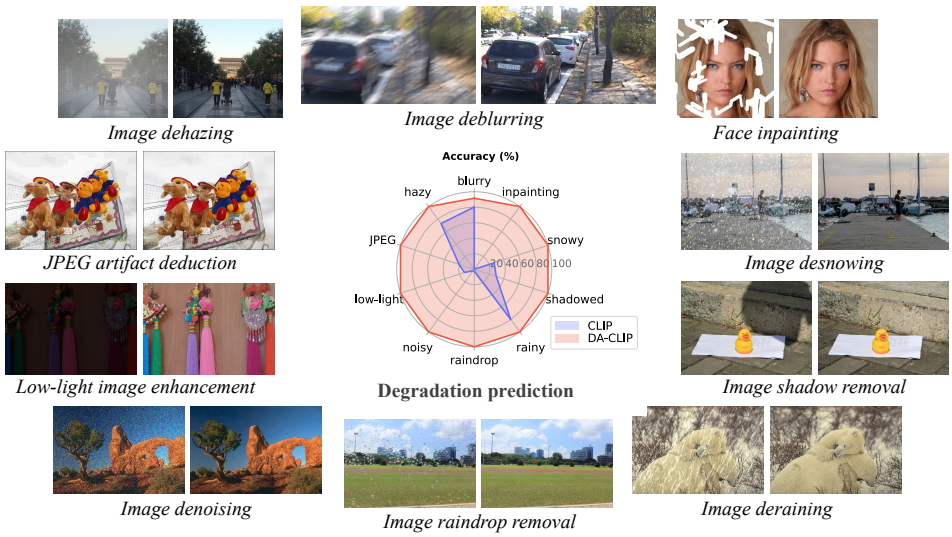


Figure 1: This paper leverages pretrained vision-language models for universal image restoration. Compared to CLIP (Radford et al., 2021), our approach precisely predicts the degradation embedding for different corrupted inputs and also outputs high-quality content features for better image restoration performance. For all examples, the right results are produced by our single unified model.

Large-scale pretrained vision-language models (VLMs) such as CLIP (Radford et al., 2021) have recently garnered significant attention, in part because of their wide-reaching usefulness on many fundamental computer vision tasks (Gu et al., 2021; Zhang et al., 2022; 2023). However, existing VLMs have so far had limited impact on low-level vision tasks such as image restoration, presumably because they do not capture the fine-grained difference between image degradation types such as "blurry" and "noisy" (Ni et al., 2023). Consequently, the existing VLMs often misalign image features to degradation texts. This is not surprising, considering that VLMs are generally trained on diverse, web-scale datasets, in contrast to most image restoration models which are trained on comparatively small datasets that are curated for a specific task without corresponding image-text pairs (Li et al., 2019; Zhang et al., 2017; Zamir et al., 2022).

Oftentimes, image restoration methods simply learn to generate images pixel-by-pixel using an ℓ_1 or ℓ_2 loss without leveraging knowledge of the degradation type. A recent line of work has, however, focused on *unified* image restoration, where the models are trained on a mixed degradation dataset and implicitly classify the type of degradation in the restoration process (Li et al., 2022a; Potlapalli et al., 2023). While the results are impressive, they are still limited to a small number of degradation types and the specific datasets that go with them. In particular, they do not make use of the vast amount of information embedded in VLMs.

In this paper, we combine large-scale pretrained vision-language models with image restoration networks and present an effective framework for universal image restoration. Specifically, aiming at addressing feature mismatching between corrupted inputs and clean captions, we propose an *Image Controller* that adapts the VLM's image encoder to output high-quality (HQ) content embeddings aligned with clean captions. Meanwhile, the controller itself also predicts a degradation embedding to match the real degradation types. This novel framework, which we call degradation-aware CLIP (DA-CLIP), incorporates the human-level knowledge from VLMs into general networks that improve image restoration performance and enable unified image restoration.

To train DA-CLIP to learn high-quality features and degradation types from low-quality (LQ) inputs, we construct a large mixed-degradation dataset for ten different image restoration tasks. Specifically, we use BLIP (Li et al., 2022b), a bootstrapped vision-language framework, to generate synthetic captions for all HQ images and then match LQ images with captions and corresponding degradation types as image-text-degradation pairs. Once trained, our DA-CLIP accurately classifies the ten different degradation types and can readily be integrated into existing restoration models, helping produce visually appealing results across the different degradations, as illustrated in Figure 1.

Our main contributions are summarised as follows:

- We present DA-CLIP to formulate universal image restoration with vision-language models. The key component is an image controller that predicts degradation and adapts the fixed CLIP image encoder to output high-quality content embeddings from corrupted inputs.
- We use cross-attention to integrate high-quality content embedding into different image restoration networks to improve their performance. Moreover, we introduce a prompt learning module to better utilize the degradation context for unified image restoration.
- We construct a mixed degradation dataset containing ten different degradation types with high-quality synthetic captions. This dataset can be used to train either DA-CLIP or a unified image restoration model.
- We demonstrate the effectiveness of DA-CLIP by applying it to image restoration models for both degradation-specific and unified image restoration. Our approach achieves highly competitive performance across all ten degradation types.

2 Background and Related Work

Image restoration Image restoration aims to recover a high-quality image from its corrupted counterpart, which is a fundamental and long-standing problem in computer vision and contains various tasks such as image denoising (Zhang et al., 2017), deraining (Ren et al., 2019), dehazing (Song et al., 2023), deblurring (Kupyn et al., 2018), etc. Most existing works focus on the *degradation-specific* task which trains multiple models for different tasks separately by directly learning the target with common pixel-based losses (e.g., ℓ_1 or ℓ_2). Recently, however, increasing attention has been

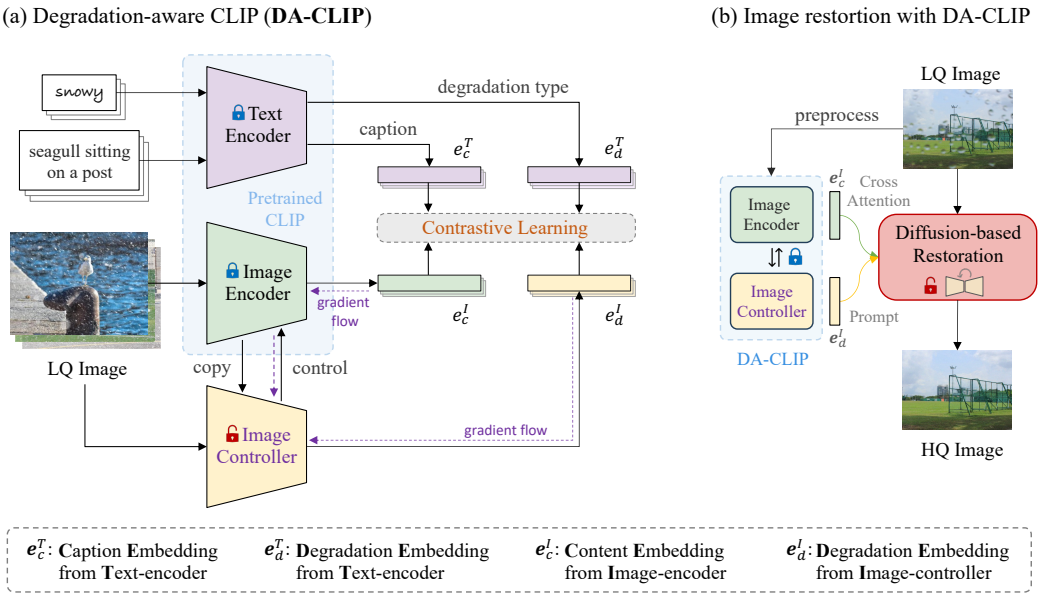


Figure 2: Overview of our method. DA-CLIP freezes both the text and image encoders of a pre-trained CLIP but learns an additional image controller with contrastive learning. This controller predicts degradation features to match real corruptions and then controls the image encoder to output high-quality content features. Once trained, DA-CLIP can be integrated into other image restoration models by simply adding a cross-attention module and a degradation feature prompting module.

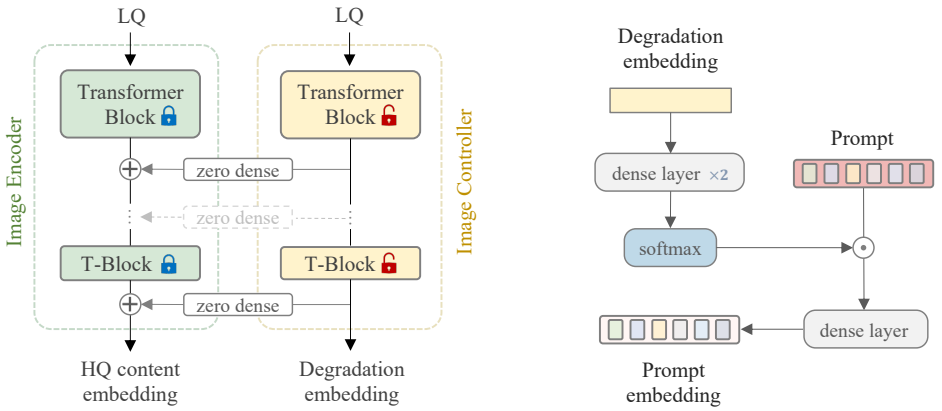
paid to *unified image restoration* where multiple image restoration tasks are handled with a single model (Li et al., 2022a; Potlapalli et al., 2023). These works achieve impressive results by using an additional encoder (Li et al., 2022a; Zhou et al., 2022) or a visual prompt module (Potlapalli et al., 2023) to implicitly cluster inputs according to the degradation type. However, they are still limited to a few image restoration tasks and do not consider auxiliary information about the degradation type.

Vision-language models (VLMs) Recent works have demonstrated the great potential of applying pretrained VLMs to improve downstream tasks with generic visual and text representations (Radford et al., 2021; Jia et al., 2021; Li et al., 2022b). A classic VLM usually consists of a text encoder and an image encoder and tries to learn aligned multimodal features from noisy image-text pairs with contrastive learning (Radford et al., 2021). BLIP (Li et al., 2022b) further proposes to remove noisy web data by bootstrapping synthetic captions. Although VLMs provide a strong capability of zero-shot and label-free classification for downstream tasks, they have so far had limited effect on image restoration due to the need for specialized and accurate terminology. A noteworthy approach for fine-tuning vision-language models is so-called *prompt learning* (Zhou et al., 2022), where the prompt's context words are represented by learnable vectors that are then optimized for the downstream task.

Text-to-image generation Text-to-image models such as stable diffusion (Rombach et al., 2022) have gained extraordinary attention from both researchers and the general public. ControlNet (Zhang & Agrawala, 2023) builds upon this work and proposes to add controls to the diffusion network to make it adapt to task-specific conditions. InstructPix2Pix (Brooks et al., 2023) further combines GPT-3 (Brown et al., 2020) with stable diffusion to perform an instruction-based image-to-image translation. Although it can generate highly realistic images, it can not be directly applied to image restoration tasks since the latter requires highly accurate reconstruction abilities.

3 Degradation-Aware CLIP

At the core of our approach is the idea of controlling a pre-trained CLIP model to output the high-quality image feature from a corrupted image while simultaneously predicting the degradation type. As summarised in Figure 2, the image content embedding $\boldsymbol{e}_c^I$ matches the clean caption embedding



(a) Controller for ViT-based image encoder

(b) Prompt with degradation embeddings

Figure 3: (a) Controlling the ViT-based image encoder with a trainable controller. (b) Learning the prompt with degradation embeddings predicted from the controller.

$\boldsymbol{e}_c^T$. Moreover, the image degradation embedding $\boldsymbol{e}_d^I$ predicted by the controller specifies the corruption type of the input, i.e. the corresponding degradation embedding $\boldsymbol{e}_d^T$ from text encoder. These features can then be integrated into other image restoration models to improve their performance.

3.1 Image Controller

The image controller is a copy of the CLIP image encoder but wrapped with a few zero-initialised connections to add controls to the encoder. It manipulates the outputs of all encoder blocks to control the prediction of the image encoder. In this paper, we use ViT (Dosovitskiy et al., 2020) as the default backbone for both the encoder and the controller. Figure 3(a) illustrates the controlling procedure, where the output of the controller consists of two parts: an image degradation embedding $\boldsymbol{e}_d^I$ and hidden controls $\boldsymbol{h}_c$. Note that the latter contains all outputs from the transformer blocks, which are subsequently added to the corresponding encoder blocks to control their predictions. The connections between the transformer blocks are simple dense neural networks with all parameters initialized to zero, which gradually influence the image encoder during training (Zhang & Agrawala, 2023). Since our training dataset is tiny compared to the web-scale datasets used in VLMs, this controlling strategy alleviates overfitting while preserving the capability of the original image encoder.

Optimising the image controller We freeze all weights of the pretrained CLIP model and only fine-tune the image controller. To make the degradation-embedding spaces discriminative and well-separated, we use a contrastive objective (Tian et al., 2020) to learn the embedding matching process. Let N denote the number of paired embeddings (from text encoder and image encoder/controller) in a training batch. The contrastive loss is defined as:

$$\mathcal{L}_{\text{con}}(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp\left(\boldsymbol{x}_i^\intercal \boldsymbol{y}_i/\tau\right)}{\sum_{j=1}^{N}\exp(\boldsymbol{x}_i^\intercal \boldsymbol{y}_j/\tau)}\right), \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are normalised vectors, and τ is a learnable temperature parameter that controls the contrastive strength. Minimising Equation 1 amounts to optimising the cosine similarity between correctly paired embeddings while enlarging the difference with other embeddings, similar to the cross-entropy loss (Radford et al., 2021). To optimise both content and degradation embeddings, we use the following joint objective:

$$\mathcal{L}_c(\omega) = \mathcal{L}_{\text{con}}(\boldsymbol{e}_c^I, \boldsymbol{e}_c^T; \omega) + \mathcal{L}_{\text{con}}(\boldsymbol{e}_d^I, \boldsymbol{e}_d^T; \omega), \tag{2}$$

where ω represents the learnable parameters of the controller. Note that all image-based embeddings (i.e., $\boldsymbol{e}_c^I$ and $\boldsymbol{e}_d^I$) are obtained from the LQ image and all text-based embeddings (i.e., $\boldsymbol{e}_c^T$ and $\boldsymbol{e}_d^T$) are from the clean caption and real degradation. Learning to align these embeddings enables DA-CLIP to predict real degradation types and HQ content features for corrupted image inputs.

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

3.2 Image Restoration with DA-CLIP

We use IR-SDE (Luo et al., 2023a) as the base framework for image restoration. It adapts a U-Net architecture similar to DDPM (Ho et al., 2020) but removes all self-attention layers. To inject clean content embeddings into the diffusion process, we introduce a cross-attention (Rombach et al., 2022) mechanism to learn semantic guidance from pre-trained VLMs. Considering the varying input sizes in image restoration tasks and the increasing cost of applying attention to high-resolution features, we only use cross-attention in the bottom blocks of the U-Net for sample efficiency.

On the other hand, the predicted degradation embeddings are useful for unified image restoration, where the aim is to process low-quality images of multiple degradation types with a single model (Li et al., 2022a). As illustrated in Figure 1, our DA-CLIP accurately classifies the degradation across different datasets and various degradation types, which is crucial for unified image restoration (Li et al., 2022a). Moreover, to make use of these degradation embeddings, we combine them with a prompt learning (Zhou et al., 2022) module to further improve the results, as shown in Figure 3(b). Given state x_t and low-quality image μ, our final diffusion network is conditioned on the time t and additional embeddings $\boldsymbol{e}_c^I$ and $\boldsymbol{e}_d^I$, as $\epsilon_\theta(x_t, \mu, t, \boldsymbol{e}_c^I, \boldsymbol{e}_d^I)$, which can be trained with either noise-matching loss (Ho et al., 2020) or maximum likelihood loss (Luo et al., 2023a;b).

Generally, we can use cross-attention to integrate content embedding into networks to improve their performance on an image restoration task. In contrast, the prompt module combined with degradation embedding specifically aims to improve the classification of the degradation type in the context of unified image restoration.

4 Dataset Construction

A crucial point of this paper is to leverage powerful vision-language models to learn multiple image degradations and extract degradation-free features for image restoration. For this purpose, we collect a large dataset with ten different image degradation types: *blurry*, *hazy*, *JPEG-compression*, *low-light*, *noisy*, *raindrop*, *rainy*, *shadowed*, *snowy*, and *inpainting*. Table 1 summarises the tasks and the number of training and testing images for each degradation type, and more details are provided in Appendix A. Example low-quality images for the ten degradations are also shown in Figure 1.

Generating image-text-degradation pairs In order to train DA-CLIP on the mixed-degradation dataset, we use the bootstrapped vision-language framework BLIP (Li et al., 2022b) to generate synthetic captions for all HQ images. Since the inputs are clean, the generated captions are assumed to be accurate and of high-quality. As illustrated in Figure 4, we are then able to construct image-text-degradation pairs by directly combining these clean captions, LQ images, and the corresponding degradation types. This dataset allows us to train either vision-language models (based on image-text-degradation pairs) or a unified image restoration framework (based on LQ-HQ image pairs).

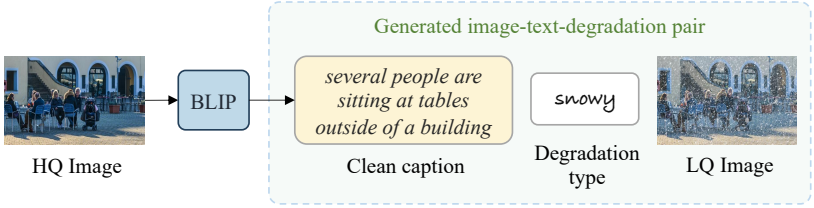


Figure 4: An example of generating the image-text-degradation tuple with BLIP. The clean caption generated from the HQ image is accurate and does not convey the degradation information.

5 EXPERIMENTS

We experimentally evaluate our method on two types of tasks: *degradation-specific* image restoration and *unified* image restoration. In the degradation-specific setting (Section 5.2), restoration models are separately trained for each of the considered degradation types. In unified image restoration (Section 5.3), a single model is instead jointly trained on all degradation types. Implementation details and additional results are provided in Appendix B and Appendix C.

5.1 MODEL EVALUATION

Our construction DA-CLIP is mainly evaluated in terms of how it affects the performance of the downstream image restoration model, which we evaluate both on multiple degradation-specific tasks and on the unified image restoration task. IR-SDE (Luo et al., 2023a) is used as our baseline model. We use the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) and Fréchet inception distance (FID) (Heusel et al., 2017) as our main metrics for perceptual evaluation, but also report PSNR and SSIM (Wang et al., 2004) for reference. Additionally, we evaluate DA-CLIP in terms of how well it can classify the ten different degradation types on the mixed dataset.

5.2 DEGRADATION-SPECIFIC IMAGE RESTORATION

We integrate our DA-CLIP into the baseline diffusion model IR-SDE and evaluate them on four degradation-specific tasks: image deraining on the Rain100H dataset (Yang et al., 2017), low-light image enhancement on LOL (Wei et al., 2018), image deblurring on the GoPro dataset (Nah et al., 2017), and image dehazing on RESIDE-6k (Qin et al., 2020). All training and testing datasets are taken from the mixed degradation dataset described in Section 4.

Comparison approaches For all tasks, we compare our method to the prevailing approaches in their respective fields such as 1) JORDER (Yang et al., 2019), PReNet (Ren et al., 2019), and MPRNet (Zamir et al., 2021) for deraining; 2) EnlightenGAN (Jiang et al., 2021), MIRNet (Zamir et al., 2020), and URetinex-Net (Wu et al., 2022) for low-light enhancement; 3) DeepDeblur (Nah et al., 2017), DeblurGAN (Kupyn et al., 2018), and DeblurGAN-v2 (Kupyn et al., 2019) for GoPro deblurring; 4) GCANet (Chen et al., 2019), GridDehazeNet (Liu et al., 2019), and DehazeFormer (Song et al., 2023) for image dehazing. We also compare with MAXIM (Tu et al., 2022), an advanced network architecture that achieves state-of-the-art performance on multiple degradation-specific tasks.

Results The quantitative results on different datasets are summarised in Table 2. Our method achieves the best perceptual results across all tasks, and even sets a new state-of-the-art performance for all metrics on the image deraining task. Compared to the baseline method IR-SDE, our approach consistently improves its results for all datasets and metrics, demonstrating that the HQ content embedding from DA-CLIP leads to better performance for downstream image restoration models. A visual comparison of our method with other approaches is also illustrated in Figure 5. Our method produces mostly clear and visually appealing results which are close to the ground truth HQ images.

5.3 UNIFIED IMAGE RESTORATION

We evaluate our method for unified image restoration on the mixed degradation dataset which contains ten different degradation types (see Section 4 for details).

Comparison approaches We compare our method with two recent approaches: AirNet (Li et al., 2022a) and PromptIR (Potlapalli et al., 2023). The former trains an extra encoder to differentiate degradation types using contrastive learning, whereas the latter employs a visual prompt module to guide the restoration. Both of them are designed specifically for unified image restoration.

Results We illustrate the comprehensive comparisons in Figure 6 and also provide the average results across all ten degradation types in Table 3. The results show that our method achieves the best perceptual results (especially in terms of FID) across the ten degradations, while still having a good distortion performance. More importantly, by simply integrating DA-CLIP into the network,

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.

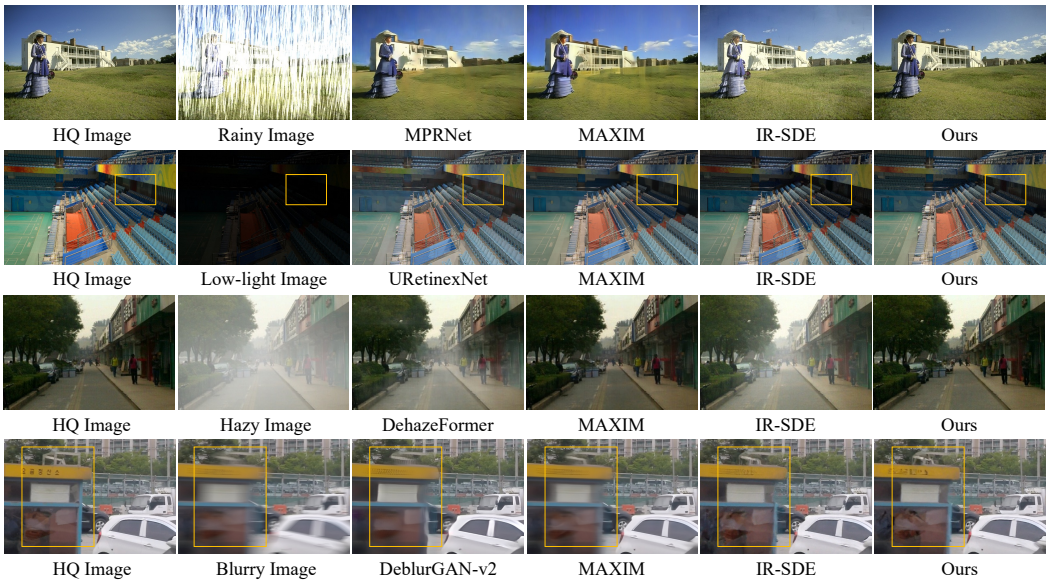


Figure 5: Comparison of our method with other approaches on 4 different *degradation-specific* tasks.

we significantly outperform the IR-SDE baseline in terms of all four metrics. A visual comparison is shown in Figure 8. On JPEG and noise removal, our method produces realistic-looking but somewhat noisy images, leading to lower distortion metrics for those particular tasks.

Moreover, we provide an extra experiment in which we integrate[1] our DA-CLIP into NAFNet (Chen et al., 2022), a relatively simple network architecture shown to achieve good image restoration performance. The results are illustrated in Table 3 and Figure 7. We observe that the modified NAFNet improves the results for all degradations and even outperforms PromptIR across all metrics. This demonstrates that DA-CLIP can be integrated with both diffusion-based and direct restoration models, further improving their performance for unified image restoration.

Finally, we evaluate DA-CLIP in terms of degradation type classification as shown in Figure 1 and Table 4 (in the Appendix). The original CLIP model achieves less than 2% accuracy in recognizing

[1]We add the prompt module to all blocks and use cross-attention in the bottom layers.

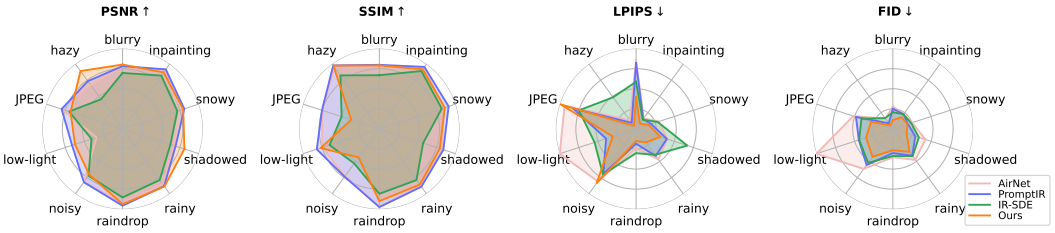


Figure 6: Comparison of our method with AirNet, PromptIR and IR-SDE for *unified* image restoration. Each radar plot reports results for the ten different degradation types, for one particular metric. For the perceptual metrics LPIPS and FID (the two rightmost plots), a lower value is better.

Table 3: Quantitative comparison between our method with other approaches on the *unified* image restoration task. We report the average results over ten different datasets.

Method	Distortion		Perceptual	
	PSNR↑	SSIM↑	LPIPS↓	FID↓
NAFNet	26.34	0.847	0.159	55.68
NAFNet + DA-CLIP	**27.22**	**0.861**	0.145	47.94
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**

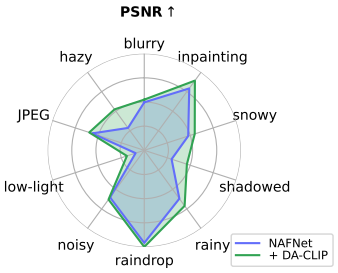


Figure 7: NAFNet with DA-CLIP for *unified* image restoration.

noisy and raindrop images, and even 0% accuracy for inpainting, which would confuse a downstream unified image restoration model. By training the controller on the mixed degradation dataset, DA-CLIP perfectly predicts all degradations except blurry, for which it achieves 91.6% accuracy.

5.4 Discussion and Analysis

We have two types of embeddings from the DA-CLIP: HQ content embedding and degradation embedding. The former is obtained from the image encoder and used for improving general image restoration performance, whereas the latter is predicted by the controller and used to classify the image degradation type for unified image restoration. To verify their effectiveness, we train the baseline model with different embedding settings. As can be observed in Figure 9a, separately applying either the degradation embedding or the HQ content embedding improves the performance. Moreover, the performance is further improved by applying both embeddings as in our DA-CLIP.

As an alternative to the DA-CLIP degradation embedding (which is predicted from the LQ image by the controller), the degradation type could be encoded directly using the CLIP text encoder. This alternative approach thus requires access to the ground truth degradation label for each LQ image. Surprisingly, we observe in Figure 9b that the two embeddings give very similar performance, meaning that the DA-CLIP controller has learned to accurately predict the degradation type.

We also compare our DA-CLIP with the approach of integrating image embeddings from the original OpenAI CLIP (Radford et al., 2021) into the baseline restoration model. For the unified image restoration task, we observe in Figure 9c that using the original CLIP embedding fails to substantially improve the performance. For the degradation-specific setting (on the image deraining task), we observe in Figure 9d that while the performance is improved with both embeddings, our DA-CLIP gives a significantly larger performance gain.

While we have demonstrated the effectiveness of our proposed DA-CLIP in various settings, it is worth acknowledging one potential limitation: increased model complexity. As can be observed in Table 5 in the Appendix, DA-CLIP significantly increases the memory requirements compared to the baseline model. The test-time computational cost is however virtually unaffected.

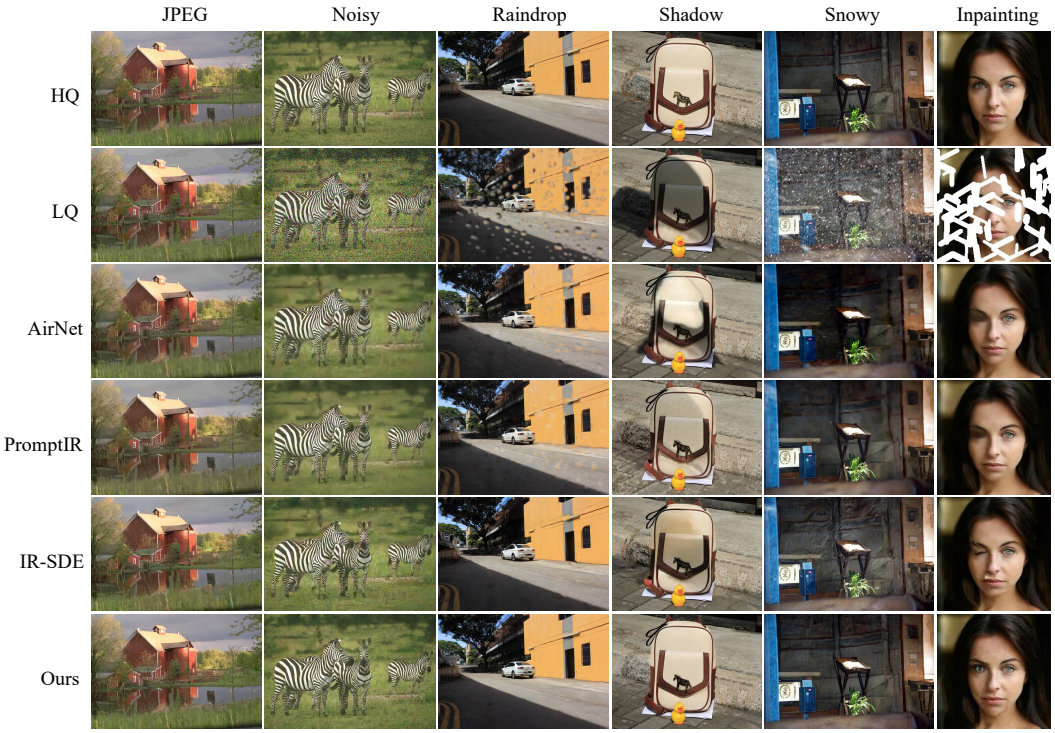
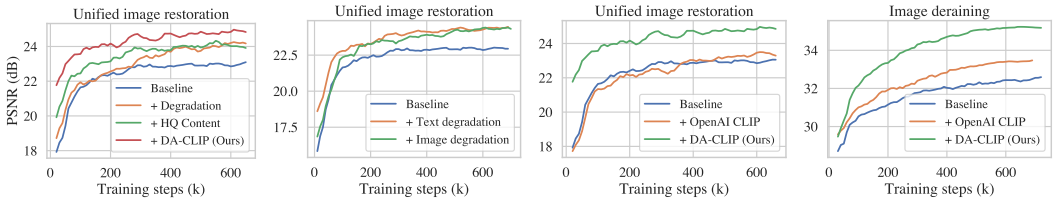


Figure 8: Comparison of our method with other approaches on the *unified* image restoration. All results in each row are produced by sending images with different degradations to the same model.



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

Figure 9: Training curves of model variations, demonstrating the effectiveness of our DA-CLIP.

6 Conclusion

This paper presents DA-CLIP to leverage large-scale pretrained vision-language models for universal image restoration. At the core of our approach is a controller that accurately predicts the degradation embeddings for low-quality images, and also controls the CLIP image encoder to output high-quality content embeddings. To train DA-CLIP, we construct a mixed degradation dataset containing generated captions for the high-quality images. DA-CLIP is then integrated into downstream image restoration models using a prompt learning module and a cross-attention mechanism. Experimental evaluation on both degradation-specific and unified image restoration tasks demonstrates that DA-CLIP consistently improves the restoration performance, across a variety of degradation types. The comprehensive analysis of DA-CLIP further verifies the effectiveness of our approach.

Future work Since the mixed degradation dataset contains a single degradation label for each image, our current model has not been trained to restore multiple degradations in the same scene. For example, the raindrop image in Figure 8 contains a shadow area, but our model only removes the raindrop degradation. Creating models which are able to fully restore such images, containing multiple degradations of different types, is an interesting direction for future work.

ACKNOWLEDGEMENTS

This research was supported by the *Wallenberg AI, Autonomous Systems and Software Program (WASP)* funded by the Knut and Alice Wallenberg Foundation; by the project *Deep Probabilistic Regression – New Models and Learning Algorithms* (contract number: 2021-04301) funded by the Swedish Research Council; and by the *Kjell & Märta Beijer Foundation*. The computations were enabled by the *Berzelius* resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 126–135, 2017. 14

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023. 3

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1375–1383. IEEE, 2019. 6

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision*, pp. 17–33. Springer, 2022. 7, 16

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2021. 2

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. 6

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020. 5

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021. 3

Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021. 6

Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8183–8192, 2018. 2, 6

Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8878–8887, 2019. 6

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17452–17462, 2022a. 2, 3, 5, 6, 16

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022b. 2, 3, 5, 17

Siyuan Li, Iago Breno Araujo, Wenqi Ren, Zhangyang Wang, Eric K Tokuda, Roberto Hirata Junior, Roberto Cesar-Junior, Jiawan Zhang, Xiaojie Guo, and Xiaochun Cao. Single image deraining: A comprehensive benchmark analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3838–3847, 2019. 2

Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7314–7323, 2019. 6

Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6):3064–3073, 2018. 14

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 14

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022. 14

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 23045–23066. PMLR, 2023a. 5, 6, 15, 16

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2023b. 5

David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the 18th IEEE International Conference on Computer Vision (ICCV)*, volume 2, pp. 416–423. IEEE, 2001. 14

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3883–3891, 2017. 6, 14

Minheng Ni, Xiaoming Li, and Wangmeng Zuo. NUWA-LIP: Language-guided image inpainting with defect-free VQGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14183–14192, 2023. 2

Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one blind image restoration. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3, 6, 16

Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2482–2491, 2018. 14

Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. FFA-Net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11908–11915, 2020. 6, 14

Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4067–4075, 2017. 14

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021. 1, 2, 3, 4, 8

Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: a better and simpler baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3937–3946, 2019. 2, 6

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. 3, 5

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 14

H Sheikh. Live image quality assessment database release 2. *http://live. ece. utexas. edu/research/quality*, 2005. 14

Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. 2, 6

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, pp. 776–794. Springer, 2020. 4

Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. NTIRE 2017 challenge on single image super-resolution: methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 114–125, 2017. 14

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5769–5780, 2022. 6

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 6, 14

Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5901–5910, 2022. 6

Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1357–1366, 2017. 6, 14

Wenhan Yang, Robby T Tan, Jiashi Feng, Zongming Guo, Shuicheng Yan, and Jiaying Liu. Joint rain detection and removal from a single image with contextualized deep networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(6):1377–1393, 2019. 6

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Proceedings of the 16th European Conference on Computer Vision*, pp. 492–511. Springer, 2020. 6

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14821–14831, 2021. 6

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5728–5739, 2022. 2

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*, 2023. 2

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26 (7):3142–3155, 2017. 2

Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3, 4

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pp. 493–510. Springer, 2022. 2

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018. 6

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3, 5

A More Details about Datasets

In this section, we give more details about the mixed degradation dataset in Section 4. We collect images for 10 different image restoration tasks, including *blurry*, *hazy*, *JPEG-compressing*, *low-light*, *noisy*, *raindrop*, *rainy*, *shadowed*, *snowy*, and *inpainting*, as shown in Figure 1. The details of these datasets are listed below:

- *Blurry*: collected from the GoPro (Nah et al., 2017) dataset containing 2103 and 1111 training and testing images, respectively.
- *Hazy*: collected from the RESIDE-6k (Qin et al., 2020) dataset which has mixed indoor and outdoor images with 6000 images for training and 1000 images for testing.
- *JPEG-compressing*: the training dataset has 3440 images collected from DIV2K (Agustsson & Timofte, 2017) and Flickr2K (Timofte et al., 2017). The testing dataset contains 29 images from LIVE1 (Sheikh, 2005). Moreover, all LQ images are synthetic data with a JPEG quality factor of 10.
- *Low-light*: collected from the LOL (Wei et al., 2018) dataset containing 485 images for training and 15 images for testing.
- *Noisy*: the training dataset is the same as that in *JPEG-compressing* but all LQ images are generated by adding Gaussian noise with noise level 50. The testing images are from CBSD68 (Martin et al., 2001) and also added that Gaussian noise.
- *Raindrop*: collected from the RainDrop (Qian et al., 2018) dataset containing 861 images for training and 58 images for testing.
- *Rainy*: collected from the Rain100H (Yang et al., 2017) dataset containing 1800 images for training and 100 images for testing.
- *Shadowed*: collection from the SRD (Qu et al., 2017) dataset containing 2680 images for training and 408 images for testing.
- *Snowy*: collected from the Snow100K-L (Liu et al., 2018) dataset. Since the original dataset is too large (100K images), we only use a subset which contains 1872 images for training and 601 images for testing.
- *Inpainting*: we use CelebaHQ as the training dataset and divide 100 images with 100 thin masks from RePaint (Lugmayr et al., 2022) for testing.

We also provide several visual examples for each task for a better understanding of the 10 degradations and datasets, as shown in Figure 10.

B Implementation Details and More Analysis

B.1 Implementation Details

The base CLIP model uses *ViT-B-32* as the backbone for its image encoder, with weights pretrained on the LAION-2B dataset (Schuhmann et al., 2022). Built upon it, we fine-tune the DA-CLIP on the mixed degradation dataset with a batch size of $3\,136$ (784×4) and learning rate 3×10^{-5}. In preprocessing, all inputs are normalized in the range $[0, 1]$ and resized to 224×224 with bicubic interpolation. We train the DA-CLIP model on four NVIDIA A100 GPUs for 50 epochs, in approximately 3 hours. For the restoration model, we use a batch size of 16 and randomly crop images to 256×256 for data augmentation. The initial learning rate is 2×10^{-4}. We use the AdamW (Loshchilov & Hutter, 2017) optimizer ($\beta_1 = 0.9$, $\beta_1 = 0.99$) with cosine decay for a total of 700K iterations. All training is done using one A100 GPU for about 5 days.

B.2 Additional Analysis of DA-CLIP

We also provide the results of directly fine-tuning CLIP on our dataset and adding a controller but without zero-initializing its dense layers in Table 4. Note that although fine-tuning CLIP achieves slightly better performance on blurry classification, it is not able to predict high-quality content

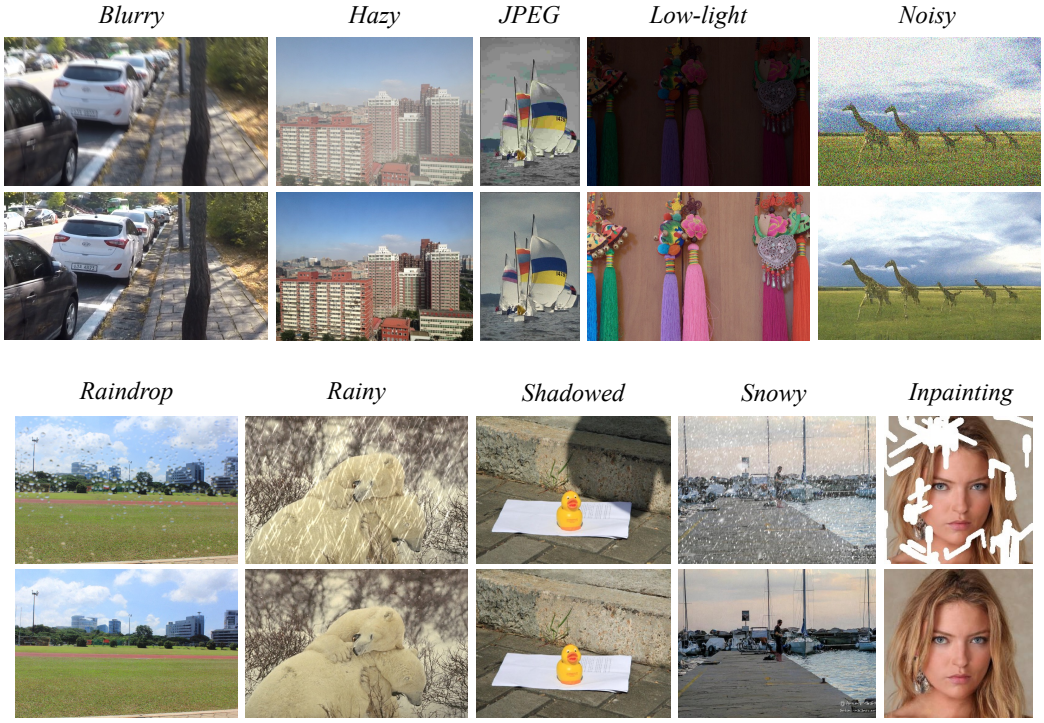


Figure 10: Example images with 10 image restoration tasks. For each task, the first row is the corrupted input and the second row is the result produced by our *unified* image restoration model.

Table 4: Accuracies of the degradation classification. The average accuracy for each method is provided in the last column. 'DA-CLIP w/o Zero' means adding a controller without zero-initialising its dense layers.

Model	Blurry	Hazy	JPEG	Low-light	Noisy	Raindrop	Rainy	Shadowed	Snowy	Inpainting	Average
OpenAI CLIP	80.3	72.9	20.7	13.3	1.5	1.7	80	29.9	27.0	0	26.7
CLIP + finetune	95.6	100	100	100	100	100	100	100	100	100	99.6
DA-CLIP w/o Zero	90.1	99	100	100	97.1	96.5	100	99.7	99	100	98.1
DA-CLIP (Ours)	91.6	100	100	100	100	100	100	100	100	100	99.2

embedding from the input, which leads to limited effects on downstream restoration models. Moreover, initializing dense layers to zero can further improve the accuracy of all datasets without adding additional costs.

B.3 Training Curves on Single Degradation Tasks

To further illustrate the effectiveness of our method, we compare the training curves between the baseline model IR-SDE (Luo et al., 2023a) and that with our DA-CLIP embeddings on four different image restoration tasks: image deraining, low-light enhancement, image deblurring, and image dehazing. The results are shown in Figure 11, from which we can see the training with our DA-CLIP obviously outperforms better than the baseline model.

B.4 Analysis about the Model Complexity

We have shown the effectiveness of applying our DA-CLIP to various image restoration models and tasks. However, it is worth acknowledging one potential limitation: the model complexity is also increased along with the DA-CLIP inference and with additional prompt modules and cross-attention layers. The comparison of model complexities is shown in Table 5.

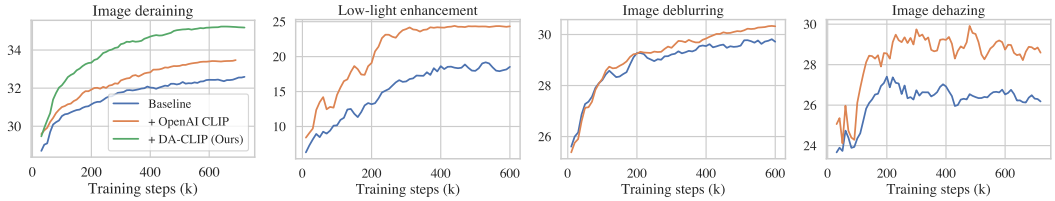


Figure 11: Ablation studies on the *degradation-specific* image restoration task.

Table 5: Comparison of the number of parameters and model computational efficiency. The flops are computed on an image with size 256×256.

METHOD	MAXIM	IR-SDE	IR-SDE + DA-CLIP
#PARAMETERS	14.1M	36.2M	48.9M + 125.2M
FLOPS	216G	117G	118G + 0.5G

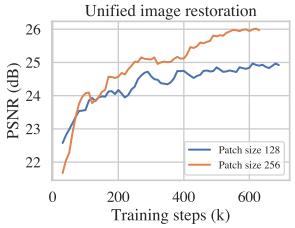


Figure 12: Comparison of training curves with different patch sizes.

B.5 ANALYSIS OF PATCH SIZES

Intuitively, large patch sizes always contain more semantic information that might be important for guiding image restoration. We explore this property by training our model with two different patch sizes on the unified image restoration task. As shown in Figure 12, increasing the patch size from 128×128 to 256×256 improves the training process, which is consistent with our conjecture.

C ADDITION EXPERIMENTAL RESULTS

In this section, we provide more results for experimental results.

C.1 DETAILED QUANTITATIVE RESULTS FOR UNIFIED IMAGE RESTORATION

We provide some examples to show the degradation prediction on specific images with different corruptions in Figure 13. In Section 5.3 we have reported the average results and radar figures for different metrics. Here we also provide a more detailed comparison between our method with other approaches for unified image restoration. The results on four metrics are shown in Table 6, Table 7, Table 8, and Table 9. Additionally, we also provide a detailed comparison between NAFNet (Chen et al., 2022) and its variant (with DA-CLIP) in all tables, which illustrates the potential of applying our DA-CLIP to other general image restoration frameworks.

C.2 ADDITIONAL VISUAL RESULTS

The additional visual results for unified image restoration are shown in Figure 14 and Figure 15. We basically compare with PromptIR (Potlapalli et al., 2023) and IR-SDE (Luo et al., 2023a) but also add the AirNet (Li et al., 2022a) generated blurry results.

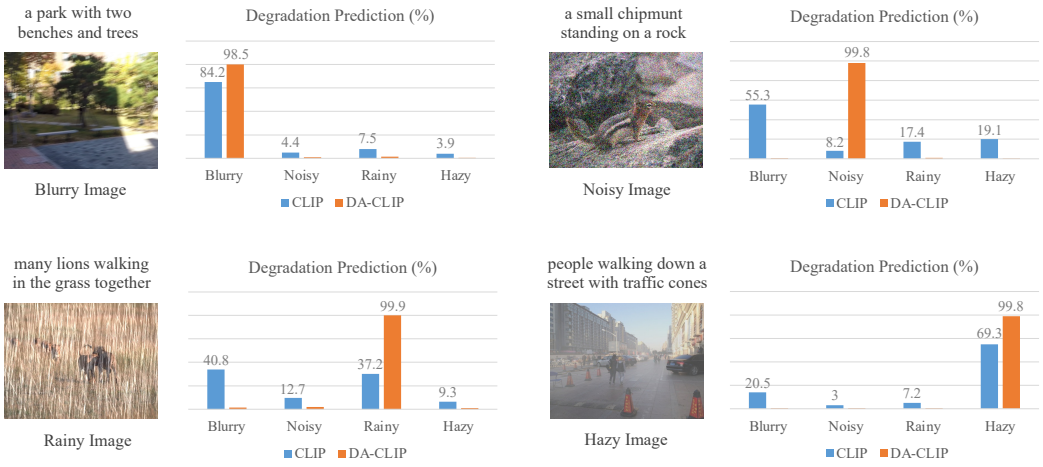


Figure 13: Comparison of CLIP and DA-CLIP for image degradation prediction. The captions above the corrupted images are generated using BLIP (Li et al., 2022b) with corresponding clean images.

Table 6: Comparison of our method with other unified image restoration approaches *on PSNR*.

Model	Blurry	Hazy	JPEG	Low-light	Noisy	Raindrop	Rainy	Shadowed	Snowy	Inpainting	Average
NAFNet	26.12	24.05	26.81	22.16	27.16	30.67	27.32	24.16	25.94	29.03	26.34
NAFNet+DA-CLIP	26.40	26.39	**27.13**	23.09	27.43	31.08	28.23	25.80	26.64	30.01	**27.22**
AirNet	26.25	23.56	26.98	14.24	27.51	30.68	28.45	23.48	24.87	30.15	25.62
PromptIR	26.50	25.19	26.95	**23.14**	**27.56**	**31.35**	29.24	24.06	**27.23**	**30.22**	27.14
IR-SDE	24.13	17.44	24.21	16.07	24.82	28.49	26.64	22.18	24.70	27.56	23.64
Ours	**27.03**	**29.53**	23.70	22.09	24.36	30.81	**29.41**	**27.27**	26.83	28.94	27.01

Table 7: Comparison of our method with other unified image restoration approaches *on SSIM*.

Model	Blurry	Hazy	JPEG	Low-light	Noisy	Raindrop	Rainy	Shadowed	Snowy	Inpainting	Average
NAFNet	0.804	0.926	0.780	0.809	0.768	0.924	0.848	0.839	0.869	0.901	0.847
NAFNet+DA-CLIP	**0.816**	**0.948**	**0.798**	0.823	**0.777**	0.928	0.869	**0.849**	0.880	0.914	**0.861**
AirNet	0.805	0.916	0.783	0.781	0.769	0.926	0.867	0.832	0.846	0.911	0.844
PromptIR	0.815	0.933	0.784	**0.829**	0.774	**0.931**	**0.876**	0.842	**0.887**	**0.918**	0.859
IR-SDE	0.730	0.832	0.615	0.719	0.640	0.822	0.808	0.667	0.828	0.876	0.754
Ours	0.810	0.931	0.532	0.796	0.579	0.882	0.854	0.811	0.854	0.894	0.794

Table 8: Comparison of our method with other unified image restoration approaches *on LPIPS*.

Model	Blurry	Hazy	JPEG	Low-light	Noisy	Raindrop	Rainy	Shadowed	Snowy	Inpainting	Average
NAFNet	0.284	0.043	0.303	0.158	**0.216**	0.082	0.180	0.138	0.096	0.085	0.159
NAFNet+DA-CLIP	0.261	0.034	0.284	0.140	0.218	0.083	0.146	0.135	0.083	0.071	27.22
AirNet	0.279	0.063	0.302	0.321	0.264	0.095	0.163	0.145	0.112	0.071	0.182
PromptIR	0.267	0.051	0.269	0.140	0.230	0.078	0.147	0.143	0.082	0.068	0.147
IR-SDE	0.198	0.168	**0.246**	0.185	0.232	0.113	0.142	0.223	0.107	0.065	0.167
Ours	**0.140**	**0.037**	0.317	**0.114**	0.272	**0.068**	**0.085**	**0.118**	**0.072**	**0.047**	**0.127**

Table 9: Comparison of our method with other unified image restoration approaches *on FID*.

Model	Blurry	Hazy	JPEG	Low-light	Noisy	Raindrop	Rainy	Shadowed	Snowy	Inpainting	Average
NAFNet	42.99	15.73	71.88	73.94	82.08	56.43	86.35	47.32	35.76	44.32	55.68
NAFNet+DA-CLIP	36.36	11.80	68.60	71.80	79.07	43.34	66.50	37.86	29.19	34.93	47.94
AirNet	41.23	21.91	78.56	154.2	93.89	52.71	72.07	64.13	36.99	32.93	64.86
PromptIR	36.5	10.85	73.02	67.15	84.51	44.48	61.88	43.24	28.29	32.69	48.26
IR-SDE	29.79	23.16	61.85	66.42	79.38	50.22	63.07	50.71	34.63	32.61	49.18
Ours	**14.13**	**5.66**	**42.05**	**52.23**	**64.71**	**38.91**	**52.78**	**25.48**	**27.26**	**25.73**	**34.89**

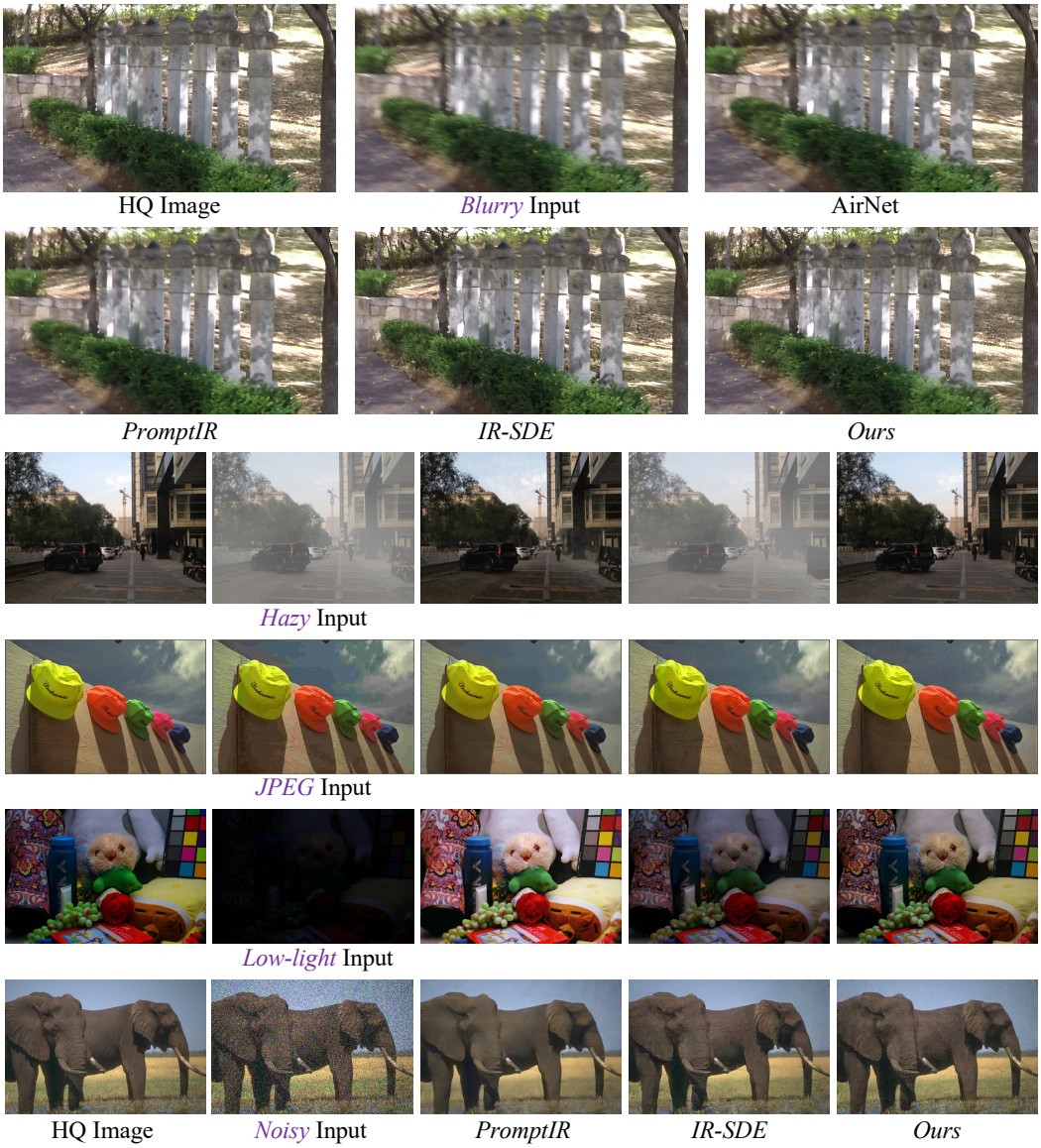


Figure 14: Comparison of our method with other approaches on the *unified* image restoration.

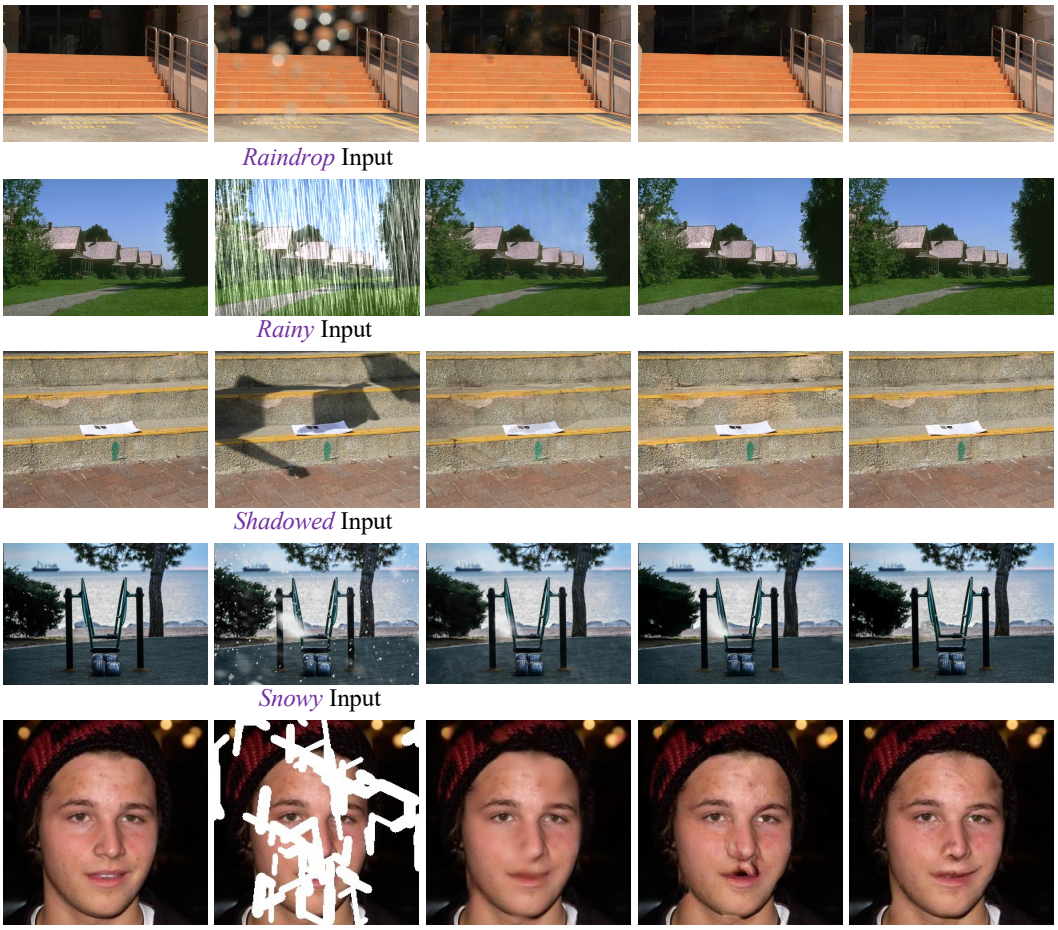


Figure 15: Comparison of our method with other approaches on the *unified* image restoration.