

Universal Image Restoration Pre-training via Masked Degradation Classification

JiaKui Hu, Zhengjian Yao, Lujia Jin, Yinghao Chen, Yanye Lu

Abstract—This study introduces a Masked Degradation Classification Pre-Training method (MaskDCPT), designed to facilitate the classification of degradation types in input images, leading to comprehensive image restoration pre-training. Unlike conventional pre-training methods, MaskDCPT uses the degradation type of the image as an extremely weak supervision, while simultaneously leveraging the image reconstruction to enhance performance and robustness. MaskDCPT includes an encoder and two decoders: the encoder extracts features from the masked low-quality input image. The classification decoder uses these features to identify the degradation type, whereas the reconstruction decoder aims to reconstruct a corresponding high-quality image. This design allows the pre-training to benefit from both masked image modeling and contrastive learning, resulting in a generalized representation suited for restoration tasks. Benefit from the straightforward yet potent MaskDCPT, the pre-trained encoder can be used to address universal image restoration and achieve outstanding performance. Implementing MaskDCPT significantly improves performance for both convolution neural networks (CNNs) and Transformers, with a minimum increase in PSNR of 3.77 dB in the 5D all-in-one restoration task and a 34.8% reduction in PIQE compared to baseline in real-world degradation scenarios. It also emerges strong generalization to previously unseen degradation types and levels. In addition, we curate and release the UIR-2.5M dataset, which includes 2.5 million paired restoration samples across 19 degradation types and over 200 degradation levels, incorporating both synthetic and real-world data. The dataset, source code, and models are available at <https://github.com/MILab-PKU/MaskDCPT>.

Index Terms—Pre-training, Degradation classification, Universal image restoration.

I. INTRODUCTION

UNIVERSAL image restoration is the process of employing a single model to transform low-quality (LQ) images affected by variable, mixed, and real-world degradation into high-quality (HQ) images. In recent work, deep learning-based methods [1, 2, 3, 4, 5, 6] have demonstrated superior performance and efficiency in solving universal image restoration compared to traditional techniques [7, 8]. The prevalent approaches employ degradation representations of LQ images as discriminative prompts for universal image restoration tasks, utilizing elements such as gradients [9], frequency [10], supplementary parameters [2], and features

JiaKui Hu, Zhengjian Yao and Yanye Lu are with Institute of Medical Technology, Peking University Health Science Center, Peking University, Beijing, China, and also with Biomedical Engineering Department, College of Future Technology, Peking University, Beijing, China, and also with National Biomedical Imaging Center, Peking University, Beijing, China.

Lujia Jin is with JIUTIAN Research, Beijing, China.

Yinghao Chen is with the College of Electronic Engineering, National University of Defense Technology, Changsha, China.

Corresponding Authors: Yanye Lu (yanye.lu@pku.edu.cn).

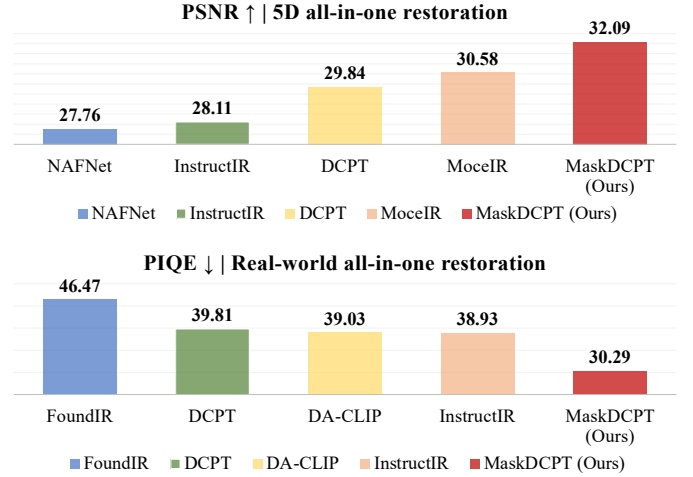


Fig. 1: MaskDCPT achieves the state-of-the-art fidelity and perception in multiple restoration tasks, including all-in-one and real-world scenarios.

compressed through large neural networks [1, 3, 4, 5, 11]. These degradation representations subsequently function as prompts for base restoration models, which are either fine-tuned or specifically trained for universal image restoration. Despite achieving high performance through the implementation of precise and effective prompts, these methods do not exploit the latent prior information inherent within the restoration models.

Pre-training methods [12, 13, 14, 15, 16, 17, 18, 19] are adept at exploiting the latent prior information inherent within the restoration models themselves. They can activate latent discriminant information within neural networks, thereby facilitating the acquisition of universal representations and rendering the pre-trained model suitable for downstream tasks. Contrastive learning [20, 21] discovers representations by maximizing agreement across multiple augmented views of the same sample using contrastive loss [22], thus obtaining features with fine-grained discriminant information [17]. Masked Image Modeling (MIM) [18, 19, 23] extends BERT's [12] success from language to vision transformers and CNNs. MIM introduces a challenging image reconstruction task through a substantially high mask ratio, which requires the model to uncover the intrinsic distribution of images. Following GPT's [13, 14] success in language generation, related methods [24] are utilized in image generation. PURE [25] also successfully used pre-trained autoregressive MLLM to adapt to real-world super-resolution. However, pre-training in image restoration [26, 27, 28] is mainly confined to single-task applications or requires carefully designed fine-tuning methods.

This suggests that current approaches do not fully arouse the universal representations provided by extensive pre-training. It is imperative to develop a pre-training framework for restoration models that can handle universal restoration tasks.

In this paper, we assert that the ability for degradation classification constitutes a frequently overlooked, yet salient, discriminative feature inherent in restoration models. We validate the effectiveness and robustness of neural networks in this capability. First, we examine the degradation classification capabilities of the classical [29, 30, 31] and all-in-one [2] image restoration architectures. Models with random initialization possess a preliminary aptitude for degradation classification, which is subsequently refined through all-in-one restoration training, thus enabling a better identification of previously unobserved degradation types. Further investigation reveals that this ability remains intact even when images are randomly masked. This observation indicates that image distribution learning based on masked modeling and degradation distribution learning based on degradation classification can coexist. Drawing upon this finding, we leverage this potential during the pre-training for universal image restoration tasks. By integrating degradation classification, restoration, and reconstruction synergistically during the pre-training phase, the model's proficiency in discerning degradation is significantly enhanced. This methodology not only maintains its efficacy in image restoration, but also fosters a more comprehensive pre-training process.

Building on these insights, we introduce a Masked Degradation Classification Pre-training (**MaskDCPT**) framework designed for universal image restoration tasks. This approach provides the model with strong prior knowledge about degradation discrimination by simultaneously pre-training three tasks: degradation classification, image reconstruction, and restoration. This enhances the model's ability to identify degradation, supporting the learning of universal restoration representations, and making the pre-trained model suitable for downstream restoration tasks. Specifically, MaskDCPT uses an encoder-decoder structure. The encoder includes an image restoration network without the restoration head. The decoder is divided into two parts: one for degradation classification and the other for image reconstruction and restoration. The encoder transforms the input image into refined latent features. The classification decoder identifies the degradation type of the input image, while the reconstruction decoder, following the MIM design, enables both reconstruction and restoration of the input image using these features. The pre-trained encoder serves as the initialization for the restoration model during fine-tuning, greatly improving restoration performance. Experimental results show that our MaskDCPT framework significantly enhances the effectiveness of various architectures in restoration tasks, including all-in-one, mixed, and real-world degradation scenarios. Moreover, to accommodate a broad spectrum of degradations present in real-world application scenarios, we curate a dataset consisting of 2.5 million samples, referred to as **UIR-2.5M**, tailored for the universal image restoration. This dataset covers 19 degradation types and over 200 degradation levels. Experiments indicate that the restoration model trained with the UIR-2.5M dataset demonstrates superior generalization when exposed to unseen degradation.

In summary, our main contributions are as follows.

- We validate that degradation classification is an inherent prior ability of restoration networks. This inherent capability is rapidly enhanced in restoration training and persists even after the input image is masked.
- We serve the degradation classification as a fundamental component of pre-training. By incorporating it with mask image modeling, we devise the MaskDCPT specifically tailored for universal image restoration.
- MaskDCPT offers substantial performance gains and can be applied to diverse architectures and tasks. Within the 5D all-in-one restoration task, MaskDCPT achieves a PSNR gain of 4.17, 4.32, 4.38, and 3.77 dB for SwinIR, NAFNet, Restormer, and PromptIR, respectively. When restoring mixed and real-world degradations, MaskDCPT provides a 34.8% reduction in PIQE to the baseline method.
- We curate and release the largest universal image restoration dataset, UIR-2.5M. The restoration model trained with UIR-2.5M shows enhanced generalization to unseen degradation types and levels.

Compared to our conference paper [32] presented at ICLR 2025, several improvements have been made in this study. The conference version segmented the pre-training into two independent stages, each assigned to degradation classification and generation capacity preservation. In this study, we accomplish the parallelization of them by integrating degradation classification with MIM, thereby enhancing training efficiency and yielding superior pre-training results. This advancement allows us to achieve the State-of-the-art (SoTA) results in 5D all-in-one restoration and mixed degradation tasks. Furthermore, we collect the dataset, UIR-2.5M, for more intricate universal image restoration. Restoration model trained with UIR-2.5M can generalize better in real-world scenarios, unseen degradation types, and levels. Finally, we expand our conference version by incorporating more references and experiments. Compared with recent methods, MaskDCPT consistently delivers superior universal image restoration performance. We further analyze how to improve MaskDCPT's pre-training performance, supported by more ablation studies.

II. RELATED WORK

A. Image Restoration

Recent advances in deep-learning based restoration models [29, 33, 34, 31, 35, 36, 30, 37, 38, 39] have consistently demonstrated superior performance and efficiency compared to traditional techniques in the realm of single task image restoration. The proposed neural networks primarily utilize convolutional neural networks (CNNs) [40] and Transformers [41]. CNNs [42, 43, 33, 34, 30] exhibit exceptional efficacy in processing localized information within images, while Transformers [29, 31, 36, 37, 39] are adept at exploiting the local self-similarity of images through the utilization of long-range dependencies. However, these methods construct specialized models tailored to individual tasks [30, 44, 39]. Consequently, a significant subset of these techniques proves insufficient to address the inherent diversities associated with image restoration [45].

Universal Image Restoration is conceived for this purpose, which requires a single model to handle various degradations. In early universal restoration approaches, distinct tasks are managed by decoupled learning [46] or employing different encoders [45] or decoder heads [47]. These approaches require the model to explicitly assess degradation types and select distinct network branches to address varied degradations. In recent developments, AirNet [1] employs MoCo [21], while IDR [48] formulates various physical degradation models to acquire degradation representations for comprehensive image restoration. PromptIR [2] integrates additional parameters via dynamic convolutions to facilitate universal image restoration without recourse to embedded features. DACLIP [4], MPerceiver [3], and DiffUIR [5] harness large external models [49, 50, 51] or generative priors to achieve improved performance and accommodate more tasks. Furthermore, VLUNet [52] has advanced the field by developing a deep unfolding network to achieve more stable restoration results. DFPIR [53] introduces degradation-related parameter perturbations. UniRestore [54] introduces considerations for task-oriented image restoration, while UniRes [55] focuses more on complex mixed degradation. These methods integrate the modulation of external parameters [2], physical models [48], human instructions [56, 6, 52], and the high-dimensional features derived from extensive neural networks [1, 3, 4, 5, 57, 58]. However, investigation into the intrinsic potential of the image restoration model and its performance ceiling has been largely overlooked.

B. Pre-training in computer vision

Pre-training is a way in which intrinsic prior are concealed in input samples and used to improve the performance in downstream tasks. In computer vision, it is divided into two schools: Contrastive Learning (CL) [20, 21] and Mask Image Modeling (MIM) [18, 19]. CL aligns features from positive pairs and uniforms the induced distribution of features in the hypersphere [59]. MIM learns to create before learning to understand [19]. However, it is difficult to extend to other architectures [23, 60, 61] and discards the decoder during downstream tasks, resulting in inconsistent representations between pre-training and fine-tuning [62]. Recently, many pre-training methods [47, 37, 63] have been proposed for restoration. Unfortunately, these methods use larger datasets to train larger models in single-degradation settings for pre-training. The existing SSL method [26] for image restoration works well in high-cost tasks but is inappropriate for low-cost tasks such as image denoising. RAM [28] pioneered the integration of MIM in the context of all-in-one image restoration. In contrast to them, MaskDCPT's focus lies on examining the influence of masks on the model's capability to discriminate degradation. Using this as a bridge, we aim to effectively merge the learning of degradation classification with the learning of image reconstruction.

III. PRELIMINARY STUDY

Beyond image reconstruction, we discern another significant and foundational ability of image restoration models: degradation classification. We verify that *(1) image restoration models*

inherently can differentiate between various degradations, (2) there is a degradation classification step in the early training of the restoration model, (3) this ability can be generalized in masked images. These findings inspire us to believe that optimizing these two intrinsic capabilities may significantly enhance the performance of the model in downstream restoration tasks.

We conduct preliminary experiments to verify these findings, using extracted output features before the restoration head and employing a k-nearest neighbor (kNN) classifier to categorize five degradation types: haze, rain, Gaussian noise, motion blur, and low-light. We randomly select 5,000 images (100 per degradation type) from the datasets: Test1200 [64] for deraining, OTS-BETA [65] for dehazing, SIDD [66] for denoising, GoPro [67] for deblurring, and LOL [68] for low-light enhancement. The images are center-cropped to maintain uniform feature dimensionality, and then the features are flattened for kNN classification. The dataset is divided into training and test sets in a 2:1 ratio, ensuring an equal distribution of degradation.

A. Inherent degradation classification ability

As presented in Figure 2 (left) at the next page, random initialized models can achieve the accuracy of the degradation classification of $52 \sim 60\%$. These models inherently possess the capability to classify degradation, highlighting that this is an intrinsic aptitude of neural networks in restoration tasks.

Drawing inspiration from self-supervised pretraining methods [18, 20, 21], it is posited that enhancing this intrinsic capability can lead to improved performance on downstream tasks. The question then arises: *In what ways can this inherent capability of restoration models be optimized?*

B. Degradation classification in restoration training

We perform an additional verification of the degradation classification capabilities of models trained in three distinct (3D) all-in-one restoration tasks (haze, rain, and Gaussian noise). It should be noted that the five target degradations for classification encompass the three types of degradation used during the training phase.

The results, as illustrated in Figure 2 (left), indicate that after 3D all-in-one training, the models exhibit an accuracy of 94% or higher in degradation classification, encompassing degradation types not previously encountered. This result suggests that the all-in-one image restoration training significantly improves the model's ability to classify degradation. Moreover, across four distinct architectures, an increase in the number of restoration training iterations corresponded to an improvement in the model's degradation classification capability. Consequently, during the training of the all-in-one restoration model, while performing restoration tasks, it simultaneously acquires the capability to discern the type of degradation present in the input image.

This experiment demonstrates that *the restoration training can optimize the model's capacity for degradation classification.* It also offers a partial clarification on the success of IPT [69].

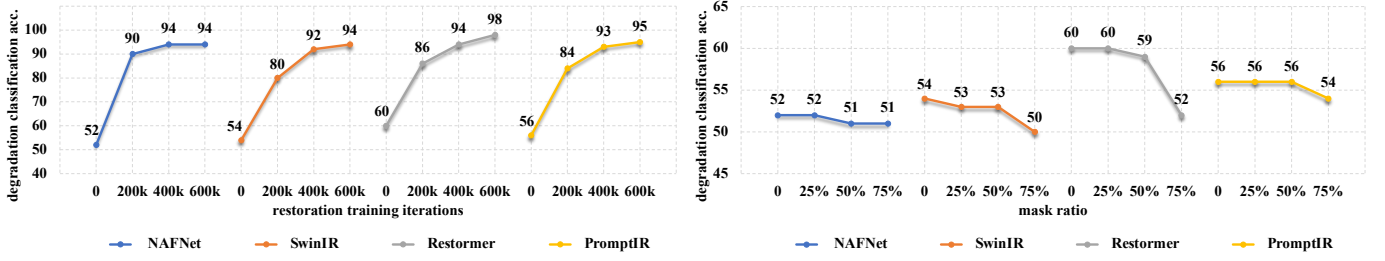


Fig. 2: Degradation classification accuracy changes with restoration training iterations (left) and image mask ratio (right). The results are averaged under five random seeds.

The direct employment of the all-in-one restoration task for pre-training serves to enhance the model’s capability in degradation classification, thereby substantially augmenting the performance on downstream tasks.

C. Degradation classification in masked images

We further investigate the robustness of the degradation classification capability of the restoration network when subjected to corrupted input images. Given its prevalence and simplicity of implementation, random masking is employed to simulate image damage. The results are illustrated in Figure 2 (right). When the mask ratio is kept below 50%, the ability to classify degradation remains largely unchanged. However, a higher mask ratio leads to a reduced classification capability.

This finding indicates that the degradation classification capability of the restoration model remains notably robust, even in scenarios where the input image is masked. Inspired by MIM [18], image reconstruction can be integrated through the application of a high-ratio mask combined with degradation classification during pre-training. This approach facilitates the learning of the image distribution and concurrently increases the degradation classification for subsequent restoration tasks, thereby enhancing the model’s restoration performance.

IV. METHOD

Based on aforementioned analysis, we propose the Masked Degradation Classification Pre-Training (MaskDCPT). We first introduce its overall pipeline and then introduce its specific components. Finally, we introduce the UIR-2.5M dataset that we collected, which includes 19 degradation types commonly seen in real life and 2.5 million images.

A. Overall Pipeline

MaskDCPT consists of an encoder that comprises restoration models [29, 31, 30, 2] without their restoration heads, and a decoder that classifies the degradation of input images based on the features of the encoder. We leverage Masked Image Modeling (MIM) to facilitate the parallel execution of both the degradation classification and image reconstruction stages. For a given input image x_{degrad} with a specified degradation D_{gt} , the image is masking according to a predetermined ratio r and then processed by the encoder to extract the encoder’s feature set F . Our decoder is equipped with two distinct heads: the reconstruction head, which serves to reconstruct the original image from F , and the degradation classification head, which

is tasked with determining the type of degradation from F . Figure 3 illustrates this overall pipeline.

B. MaskDCPT

In this section, we describe the detailed training process of MaskDCPT in components.

1) *Masked Encoder*: Given a low quality image x_{lq} , we randomly mask the degraded images (in patch size 16×16) with a mask ratio r . r is 50% by default.

$$\bar{x}_{lq} = M \odot x_{lq}, \quad (1)$$

where M is the mask map and $\odot$ is the Hadamard product.

To achieve a more effective degradation classification, it is crucial to extract features from deeper layers that contain richer high-level semantic information [70]. However, image restoration models typically adhere to the residual learning design concept [71]. The sole reliance on features from the deepest layer for the loss function calculation may result in gradient vanishing in the shallower layers, due to the loss of the encoder’s long residual connections [71] during feature extraction. To achieve a balance, the features are extracted from each block in the latter part of the encoder. We define these extracted features as $\{F_i\}$, $i \in ([\frac{l}{2}] + 1, \dots, l)$, where l is the number of blocks in the network, and $[\cdot]$ is the integer symbol.

$$\{F_i\} = \text{Encoder}(\bar{x}_{lq}). \quad (2)$$

When local operators, such as convolution, are used to process input images, these operators integrate surrounding data through the process. This causes the introduction of new signals where the signals should have been masked, leading to information aliasing. To reduce this aliasing [72], submanifold convolution [73] is employed in pretraining.

Discussion. The existing MIM-based restoration pretraining method RAM [28] uses pixel-level mask modeling, which improves the model’s ability to capture the distribution of individual pixels. However, this focus on pixel-level details might cause a lack of capturing local information within images. Ablation experiments IX show that, when applying the same fine-tuning strategy, our patch-level mask pretraining outperforms RAM’s pixel-level masking approach.

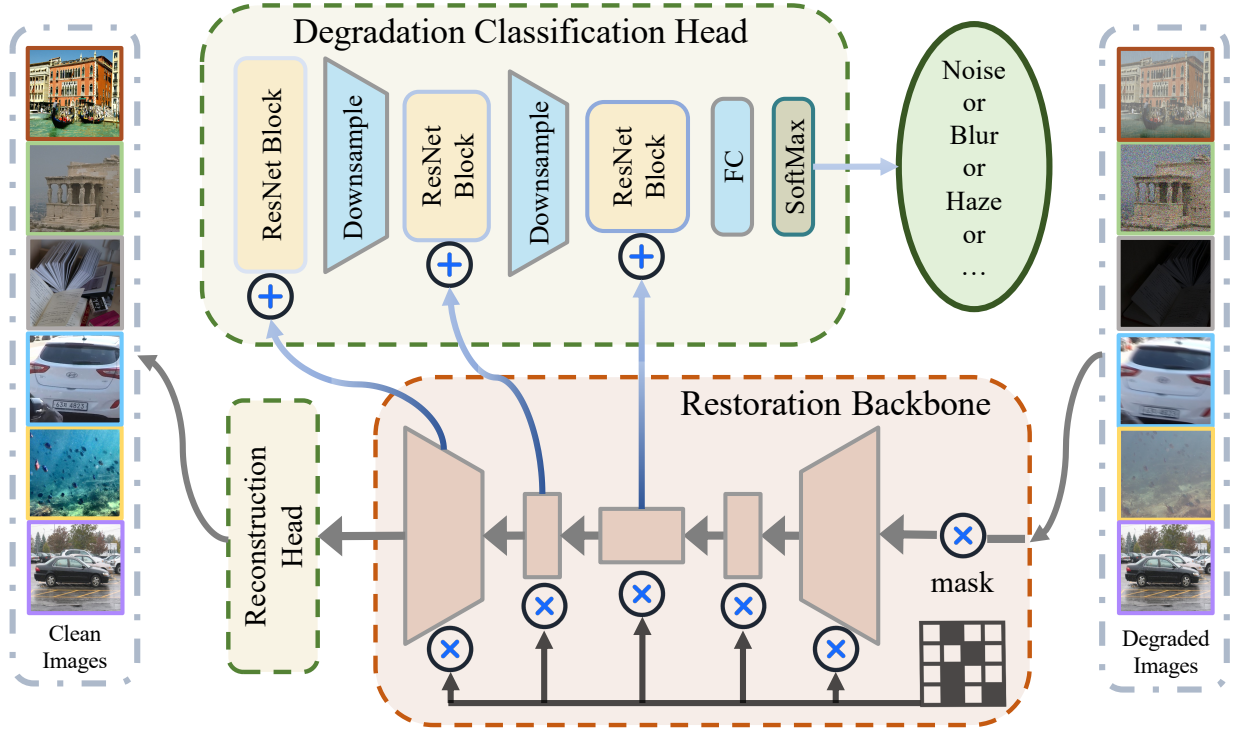


Fig. 3: MaskDCPT's overall pipeline. First, MaskDCPT receives degraded images and implements a random patch-level masking to them. Subsequently, the restoration backbone processes the masked images. Throughout this phase, network features are masked to impede the local information leakage. These masked features are then directed towards the reconstruction head for the image restoration as well as the degradation classification head for the degradation classification. After MaskDCPT, the encoder is fine-tuned for downstream restoration tasks.

2) *Classification Decoder*: After extracting masked multi-level features, we feed $\{F_i\}$ into the degradation classification decoder (**DegCls-Dec**) of the lightweight decoder to classify the degradation of the input images. The details of the decoder architecture are shown in Figure 3. To better aggregate the extracted features, it is necessary to scale up to the features $\{F_i\}$. The scaling coefficient $\{\omega_i\}$ is learnable. Then, the scaled feature $F'_i = \omega_i F_i$ is plugged into the i -th block in ResNet18 to classify the degradation. For stabilizing the training process, we replace the normalization layers in the decoder from BatchNorm to LayerNorm.

$$\hat{D}_{gt} = \text{DegCls-Dec}(\{F'_i\}). \quad (3)$$

It is crucial to note that the challenge of obtaining image restoration data [74] results in an imbalance in the number of data sets that represent different types of degradation. For example, the deraining dataset Rain200L [75] comprises only 200 images, whereas the dehazing dataset RESIDE [65] encompasses 72,135 images. This imbalance poses a significant long-tail challenge in classifying degradation. To address this issue, we employ Focal Loss [76] as the loss function for long-tail degradation classification.

$$L_{cls} = \text{Focal Loss}(D_{gt}, \hat{D}_{gt}). \quad (4)$$

3) *Reconstruction Decoder*: Another MaskDCPT's task is to enable the restoration model to learn clean image distributions by reconstructing masked images. The reconstruction decoder

(**Recon-Dec**) in the decoder allows the encoder's feature F_l to reconstruct $\hat{x}_{gt}$, as shown in Figure 3. The overall loss function of MaskDCPT is as follows:

$$\begin{aligned} L_{total} &= \alpha L_{pix} + L_{cls} \\ &= \alpha \|x_{gt} - \hat{x}_{gt}\|_1 + \text{Focal Loss}(D_{gt}, \hat{D}_{gt}), \end{aligned} \quad (5)$$

where α is 1 by default, and $\hat{x}_{gt} = \text{Recon-Dec}(F_l)$.

Discussion. Eq. 5 performs the simultaneous execution of three tasks: degradation classification, image reconstruction, and image restoration. This is because L_{pix} analyzes both the masked and unmasked regions. The former facilitates image reconstruction from unmasked regions, while the latter aids in the restoration of unblemished images from unmasked areas. As image restoration still enhances the degradation classification, it is posited that the losses in Eq. 5 are mutually reinforced, thus fostering an expedited and more robust pre-training.

Furthermore, within masking, L_{cls} in Eq. 5 acts as a bridge between MIM and CL. Unlike the ill-posed property of image restoration, degradation classification has a clear and well-defined objective. Under the training objective defined by Eq. 5, the model learns to extract information from partially masked inputs. When images undergo the same degradation but are masked differently, they should still be classified into the same degradation category, indicating a convergence of their learned representations. In contrast, images with different degradations, even if masked by the same mask, should be classified into

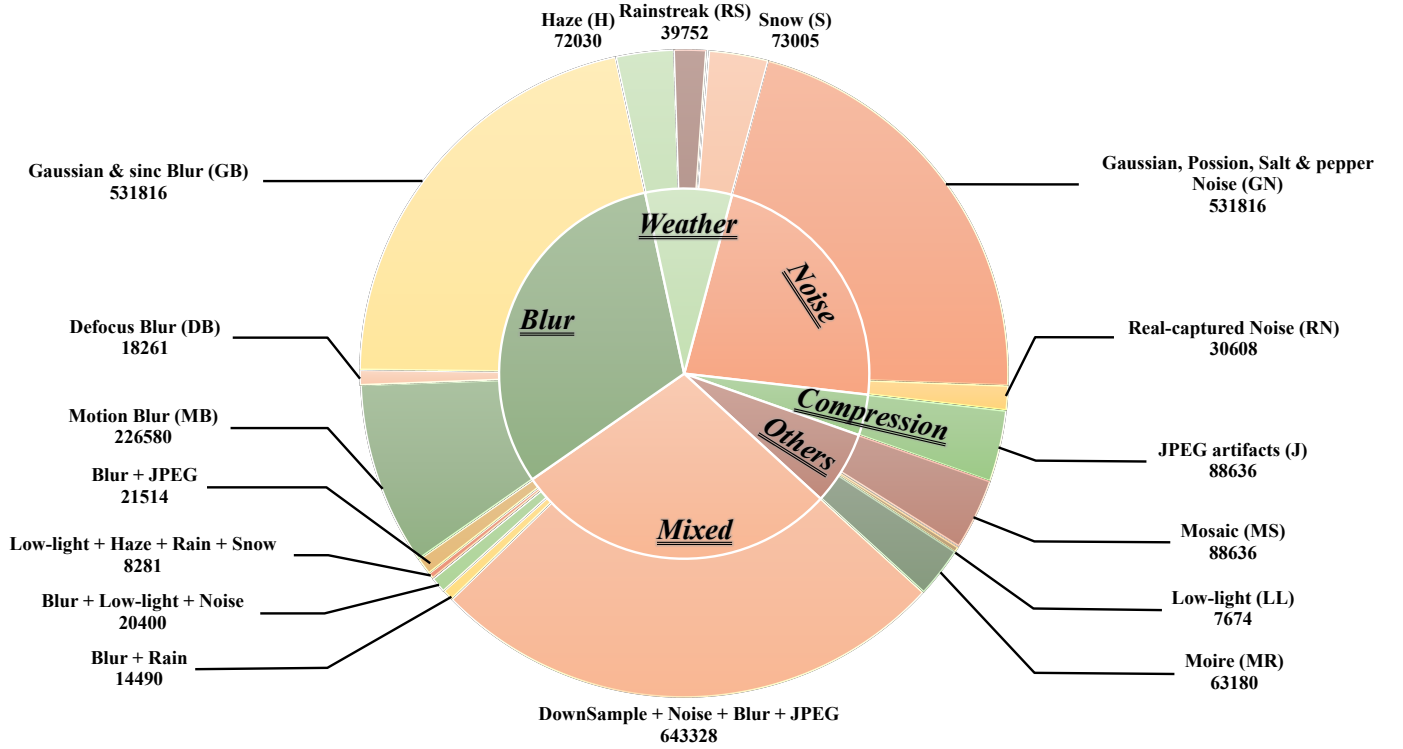


Fig. 4: The structure of the UIR-2.5M dataset. It consists of two principal categories, namely single and mixed. Single degradation types contain: blur, weather, noise, compression, and others (suboptimal imaging conditions). Mixed degradation tasks comprise combined distortions resulting from adverse weather, JPEG artifacts, motion blur, and low-light. Our dataset encompasses both synthetic and real-world data instances. The comprehensive dataset includes 2.5 million samples. The detailed distribution of the UIR-2.5M dataset is presented in the Appendix.

different degradation categories, reflecting divergent features. This behavior aligns with the core principles of CL.

C. Data collection

To address a comprehensive range of degradations encountered in real-world scenarios, an ideal approach would involve training on a large dataset that includes various degradations and features images rich in texture detail. However, since degraded images and their perfectly registered clean images cannot coexist in real environments, it is a huge challenge to construct a paired general image restoration dataset from real-world data. While generating simulated data is a relatively straightforward task and allows for the creation of complex mixed degradations not easily captured in real-world scenarios, synthetic datasets lack the diversity and realism to effectively train models capable of generalizing to demanding real-world environments. A practical strategy involves curating and filtering existing datasets, followed by preprocessing them into a standardized format conducive to research applications. Therefore, we carefully selected the available training datasets to ensure maximum coverage of different types of degradation and image textures. Table 4 provides a summary of our curated real and synthetic training datasets, categorized by degradation.

Following the aforementioned operations, a comprehensive collection of 2,482,988 pairs of universal image restoration datasets designated as **UIR-2.5M** has been assembled, encompassing *single* (1,774,975) and *mixed* (708,013) segments.

To enhance applicability in practice, it is noted that in both segments, a proportion of 3% of the data are sourced from the real-world. Fields such as face, remote sensing, medical imaging, and document remain unexplored and are thus earmarked for future work in the collation of image restoration data within those specific sub-fields. Additionally, local degradation, such as reflection, flare, and incompleteness, has yet to be addressed, with plans to focus on these challenges in future work.

V. EXPERIMENTS AND RESULTS

Our evaluation of MaskDCPT encompasses three distinct scenarios: **(1) All-in-one.** We fine-tune a single model after MaskDCPT to facilitate image restoration across various degradations, assessing its performance on both 5D all-in-one and 12D tasks. **(2) Single-task.** Following IDR [48], we assess the performance of all-in-one trained models in unseen or real-world degradations without fine-tuning. To elucidate the impact of MaskDCPT on single-task pretraining, we present the fine-tuning results of MaskDCPT pretrained models within particular single-task contexts. **(3) Mixed degradation.** We perform an evaluation of the fine-tuned model under mixed degradation conditions to determine the suitability of MaskDCPT to restore complex degraded images with mixed degradation.

Metrics. Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Metric (SSIM) within the sRGB color space

Method	Dehazing	Deraining	Denoising	Deblurring	Low-Light
AirNet [1]	21.04 / 0.884 / 0.077 / 62.52	32.98 / 0.951 / 0.058 / 50.12	30.91 / 0.882 / 0.102 / 78.12	24.35 / 0.781 / 0.189 / 66.13	18.18 / 0.735 / 0.122 / 116.9
IDR [48]	25.24 / 0.943 / 0.052 / 33.25	35.63 / 0.965 / 0.043 / 45.62	31.60 / 0.887 / 0.092 / 66.24	27.87 / 0.846 / 0.178 / 40.83	21.34 / 0.826 / 0.108 / 100.6
AdaIR [57]	30.25 / 0.981 / 0.013 / 13.11	37.86 / 0.981 / 0.014 / 13.75	31.30 / 0.892 / 0.110 / 46.64	28.11 / 0.864 / 0.189 / 19.59	22.94 / 0.894 / 0.120 / 52.41
DA-CLIP [4]	29.78 / 0.968 / 0.014 / 15.26	35.65 / 0.962 / 0.022 / 22.24	30.93 / 0.885 / 0.089 / 54.12	27.31 / 0.838 / 0.143 / 23.34	21.66 / 0.828 / 0.095 / 55.81
RCOT [77]	30.26 / 0.971 / 0.016 / 16.74	36.88 / 0.975 / 0.024 / 19.67	31.05 / 0.882 / 0.099 / 62.12	28.12 / 0.862 / 0.155 / 21.56	22.76 / 0.830 / 0.097 / 61.24
DA-RCOT [78]	30.96 / 0.975 / 0.008 / 10.62	37.87 / 0.980 / 0.012 / 12.20	31.23 / 0.888 / 0.082 / 37.65	28.68 / 0.872 / 0.135 / 12.39	23.25 / 0.836 / 0.084 / 47.23
MoceIR [79]	30.72 / 0.979 / 0.013 / 13.28	38.01 / 0.982 / 0.014 / 13.63	31.34 / 0.893 / 0.103 / 42.93	30.04 / 0.901 / 0.143 / 15.11	23.00 / 0.902 / 0.118 / 49.77
DFPIR [53]	31.23 / 0.982 / 0.013 / 13.48	37.56 / 0.979 / 0.016 / 14.71	31.26 / 0.892 / 0.091 / 38.75	28.79 / 0.879 / 0.164 / 17.07	23.79 / 0.895 / 0.122 / 56.35
SwinIR [29]	21.50 / 0.891 / 0.069 / 82.13	30.78 / 0.923 / 0.081 / 64.38	30.59 / 0.868 / 0.122 / 79.08	24.52 / 0.773 / 0.288 / 56.21	17.81 / 0.723 / 0.159 / 146.2
+ RAM [28]	28.45 / 0.975 / 0.021 / 10.19	26.09 / 0.875 / 0.209 / 92.90	31.06 / 0.888 / 0.110 / 39.95	26.88 / 0.823 / 0.249 / 36.31	21.55 / 0.876 / 0.156 / 89.10
+ DCPT [32]	28.68 / 0.977 / 0.019 / 8.93	35.70 / 0.975 / 0.022 / 12.10	31.16 / 0.890 / 0.113 / 40.00	26.42 / 0.810 / 0.270 / 37.17	20.38 / 0.836 / 0.154 / 68.46
+ MaskDCPT (Ours)	29.29 / 0.981 / 0.015 / 5.88	37.16 / 0.979 / 0.014 / 7.60	31.13 / 0.890 / 0.080 / 27.41	26.53 / 0.808 / 0.218 / 29.23	21.94 / 0.905 / 0.111 / 55.69
NAFNet [30]	25.23 / 0.939 / 0.053 / 32.68	35.56 / 0.967 / 0.050 / 43.57	31.02 / 0.883 / 0.139 / 49.57	26.53 / 0.808 / 0.206 / 49.12	20.49 / 0.809 / 0.141 / 127.9
+ DCPT [32]	29.47 / 0.976 / 0.015 / 4.26	35.68 / 0.973 / 0.021 / 12.73	31.31 / 0.886 / 0.106 / 41.88	29.22 / 0.886 / 0.153 / 15.54	23.52 / 0.855 / 0.113 / 44.57
+ MaskDCPT (Ours)	31.40 / 0.978 / 0.012 / 3.39	39.92 / 0.986 / 0.008 / 4.21	31.41 / 0.894 / 0.076 / 26.55	31.40 / 0.920 / 0.092 / 7.61	26.31 / 0.888 / 0.071 / 25.88
Restormer [31]	24.09 / 0.927 / 0.067 / 43.62	34.81 / 0.962 / 0.050 / 51.69	31.49 / 0.884 / 0.108 / 40.79	27.22 / 0.829 / 0.191 / 32.02	20.41 / 0.806 / 0.144 / 123.1
+ DCPT [32]	29.19 / 0.976 / 0.018 / 6.47	36.62 / 0.977 / 0.019 / 11.65	31.20 / 0.890 / 0.105 / 44.73	28.58 / 0.875 / 0.170 / 18.31	23.26 / 0.842 / 0.120 / 59.28
+ MaskDCPT (Ours)	32.67 / 0.985 / 0.010 / 3.12	39.27 / 0.985 / 0.009 / 4.87	31.29 / 0.892 / 0.076 / 26.50	30.58 / 0.910 / 0.102 / 11.12	26.11 / 0.879 / 0.076 / 30.33
PromptIR [2]	25.20 / 0.931 / 0.034 / 28.13	35.94 / 0.964 / 0.049 / 40.42	31.17 / 0.882 / 0.120 / 43.71	27.32 / 0.842 / 0.133 / 36.29	20.94 / 0.799 / 0.148 / 118.3
+ RAM [28]	29.63 / 0.975 / 0.014 / 5.64	28.11 / 0.888 / 0.178 / 79.38	31.08 / 0.889 / 0.109 / 42.79	28.00 / 0.862 / 0.183 / 18.36	24.45 / 0.907 / 0.120 / 51.45
+ DCPT [32]	30.93 / 0.982 / 0.012 / 3.89	37.18 / 0.979 / 0.016 / 9.75	31.27 / 0.891 / 0.110 / 46.10	28.86 / 0.880 / 0.164 / 17.61	23.09 / 0.840 / 0.128 / 60.42
+ MaskDCPT (Ours)	32.71 / 0.985 / 0.009 / 3.07	39.12 / 0.985 / 0.009 / 4.94	31.30 / 0.892 / 0.079 / 27.67	29.99 / 0.900 / 0.111 / 11.81	26.30 / 0.881 / 0.078 / 29.53

TABLE I: **5D all-in-one image restoration results** in terms of PSNR $\uparrow$ / SSIM $\uparrow$ / LPIPS $\downarrow$ / FID $\downarrow$. Classic restoration models pre-trained with MaskDCPT outperform the methods that require all-in-one specific training and architecture. All methods are trained on widely used 5D all-in-one restoration dataset following IDR [48] to ensure fair comparison.

are employed to quantify image distortions. In addition, Learned Perceptual Image Patch Similarity (LPIPS) and Fréchet Inception Distance (FID) are utilized as perceptual metrics. For test sets lacking reference HQs, the Perception-based Image Quality Evaluator (PIQE) and the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) serve as evaluation metrics. We use the pyiqa¹ to calculate them.

A. All-in-one image restoration

We first assess the performance gain of MaskDCPT on different architectures in all-in-one image restoration.

5D all-in-one dataset. To facilitate a fair comparison, a subset of UIR-2.5M was meticulously crafted for the 5D all-in-one task. This subset, termed as the *UIR-2.5M-5D* subset, comprises the following: Rain200L, consisting of 200 training images for the purpose of deraining; RESIDE, which includes 72,135 training images alongside 500 test images (SOTS) designated for dehazing; BSD400 and WED, collectively offering 5,144 training images for Gaussian denoising; GoPro, featuring 2,103 training images and 1,111 test images intended for single image motion deblurring; and LOL, which provides 485 training images accompanied by 15 test images for the low-light enhancement.

12D all-in-one dataset. For the 12D all-in-one task, we use the UIR-2.5M-single for training. In order to comprehensively assess the effectiveness of the restoration model across diverse simulated and real-world conditions, we employ the following datasets for evaluation: MSPFN 5 sets (Rain100L, Rain100H, Test100, Test1200, Test2800), SynRain-13k for deraining; SOTS for dehazing; Snow100K-L for desnowing; RainDS and RainDrop for raindrop removal; LoL v1, LoL

v2, LSRW for low-light enhancement; GoPro, HIDE, REDS for deblurring; DPDD for defocus deblurring; Urban100 for Gaussian denoising, deblurring, and demosaicing; SIDD for real-world captured denoising; and RDNet for demore.

Implementation details. During MaskDCPT, image restoration models are trained by AdamW optimizer with zero weight decay for 100k iters with batch-size 16 on 256×256 image patches on 4 NVIDIA L40S GPUs. Due to the heterogeneous encoder-decoder design, we employ different learning rates for the encoder and decoder. The learning rate is set to 3×10^{-4} for the encoder and to 1×10^{-4} for the decoder. The learning rate does not alter during MaskDCPT. After MaskDCPT, the encoder is used to initialize the image restoration models.

Dataset sampler. For degradation with fewer training data, we use repeat sampler technology to ensure that there are enough training pairs for each degradation. For 5D all-in-one image restoration, the repetition ratio is [1H, 300RS, 15GN, 5MB, 60LL]. For 12D all-in-one image restoration, the repetition ratio is [4GN, 4RN, 1MB, 20DB, 1GB, 4J, 5H, 8RS, 180RD, 5S, 4MS, 30LL, 6MR]. The above abbreviations used to represent degradation can be found in the “abbrev.” in Figure 4.

5D all-in-one image restoration results are reported in Table I and Figure 5. (1) The models pre-trained using MaskDCPT exhibit superior performance compared to the specifically designed all-in-one image restoration architecture across most tasks. In the low-light enhancement task, MaskDCPT-NAFNet achieves an improvement of 2.52 dB over DFPIR and 3.37 dB over AdaIR in terms of PSNR. Additionally, it outperforms IDR [48] by 3.53 dB and DA-RCOT [78] by 2.72 dB in the motion deblurring task. (2) Regarding the multi-stage training method (IDR), MaskDCPT consistently demonstrates performance improvements. When Restormer [31] serves as the baseline, the performance gain achieved by IDR is confined

¹<https://github.com/chaofengc/IQA-PyTorch>



Fig. 5: *Visual comparison on 5D all-in-one image restoration datasets. Zoom in for best view.*

to 0.74 dB relative to its base method, whereas MaskDCPT provides an average performance gain of 4.38 dB. (3) MaskDCPT exhibits adaptability to a wide range of architectures. Observations indicate that regardless of whether the network employs a CNN or Transformer architecture, and whether it follows a linear [29] or UNet-like structure [30, 31, 2], MaskDCPT consistently delivers an average performance enhancement of 3.77 dB and above in the 5D all-in-one image restoration task. (4) MaskDCPT demonstrates superiority over existing pre-training methods. Compared to PromptIR [2]

models pre-trained using the MaskDCPT and RAM [28] frameworks, those pre-trained with MaskDCPT show significant performance improvements. Specifically, in the dehazing task, MaskDCPT-PromptIR surpasses RAM-PromptIR by 3.08 dB and DCPT-PromptIR by 1.78 dB, respectively.

12D all-in-one image restoration results. Furthermore, the degradation types are scaled up to 12 to determine the efficacy of MaskDCPT in the presence of a greater number of degradation types. Following DACLIP [4] and InstructIR [56], we use NAFNet as the basic restoration model due to its

Type	Method	<i>Deraining</i>		Method	<i>Dehazing</i>		Method	<i>Desnowing</i>	
		PSNR / SSIM $\uparrow$	LPIPS / FID $\downarrow$		PSNR / SSIM $\uparrow$	LPIPS / FID $\downarrow$		PSNR / SSIM $\uparrow$	LPIPS / FID $\downarrow$
Task Specific	DGUNet	30.99 / 0.913	0.095 / 32.94	Dehamer	34.93 / 0.989	0.008 / 18.11	SFNet	30.21 / 0.907	0.086 / 3.34
	Restormer	31.76 / 0.924	0.082 / 28.89	MB-Talor	37.54 / 0.993	0.005 / 2.545	Focal-Net	29.97 / 0.903	0.090 / 3.43
All in One	InsturctIR [56]	29.12 / 0.891	0.100 / 34.98	InsturctIR [56]	26.90 / 0.952	0.178 / 44.03	InsturctIR [56]	28.20 / 0.858	0.108 / 4.83
	DA-CLIP [4]	29.95 / 0.902	0.071 / 24.90	DA-CLIP [4]	28.19 / 0.965	0.069 / 8.12	DA-CLIP [4]	28.26 / 0.868	0.088 / 2.10
	UniRestore [54]	27.87 / 0.864	0.125 / 39.53	UniRestore [54]	27.91 / 0.904	0.093 / 10.35	UniRestore [54]	28.78 / 0.863	0.099 / 3.89
	FoundIR [80]	29.72 / 0.890	0.094 / 33.25	FoundIR [80]	30.12 / 0.974	0.013 / 4.94	FoundIR [80]	29.64 / 0.895	0.082 / 3.36
	DCPT [32]	30.50 / 0.909	0.077 / 27.25	DCPT [32]	30.40 / 0.976	0.015 / 4.29	DCPT [32]	30.42 / 0.907	0.089 / 3.08
	MaskDCPT (Ours)	31.93 / 0.922	0.056 / 20.46	MaskDCPT (Ours)	31.82 / 0.982	0.011 / 3.19	MaskDCPT (Ours)	30.51 / 0.909	0.073 / 1.84
Type	Method	<i>Raindrop</i>		Method	<i>Low-light Enhance</i>		Method	<i>Motion Deblur</i>	
		PSNR / SSIM $\uparrow$	LPIPS / FID $\downarrow$		PSNR / SSIM $\uparrow$	LPIPS / FID $\downarrow$		PSNR / SSIM $\uparrow$	LPIPS / FID $\downarrow$
Task Specific	IDT	24.55 / 0.803	0.198 / 56.04	GLARE	19.57 / 0.766	0.183 / 50.12	Stripformer	30.43 / 0.902	0.119 / 8.80
	UDR-S2Former	27.33 / 0.815	0.340 / 44.38	LLFlow-SKF	22.60 / 0.660	0.194 / 58.53	DiffIR	30.53 / 0.898	0.128 / 9.76
All in One	InsturctIR [56]	21.19 / 0.761	0.275 / 109.5	InsturctIR [56]	21.42 / 0.752	0.208 / 55.69	InsturctIR [56]	27.85 / 0.847	0.194 / 15.43
	DA-CLIP [4]	24.03 / 0.772	0.180 / 54.91	DA-CLIP [4]	19.91 / 0.709	0.198 / 52.03	DA-CLIP [4]	26.25 / 0.822	0.177 / 15.45
	UniRestore [54]	20.57 / 0.721	0.333 / 113.2	UniRestore [54]	9.55 / 0.276	0.493 / 113.4	UniRestore [54]	26.29 / 0.807	0.194 / 20.95
	FoundIR [80]	21.10 / 0.757	0.275 / 107.3	FoundIR [80]	19.67 / 0.688	0.234 / 50.77	FoundIR [80]	27.10 / 0.827	0.169 / 17.15
	DCPT [32]	20.32 / 0.751	0.253 / 108.2	DCPT [32]	19.54 / 0.646	0.262 / 59.14	DCPT [32]	27.68 / 0.856	0.199 / 16.46
	MaskDCPT (Ours)	27.57 / 0.838	0.124 / 26.83	MaskDCPT (Ours)	24.35 / 0.794	0.168 / 34.83	MaskDCPT (Ours)	29.83 / 0.884	0.127 / 8.68
Type	Method	<i>Defocus Deblur</i>		Method	<i>JPEG Removal</i>		Method	<i>Real Denoising</i>	
		PSNR / SSIM $\uparrow$	LPIPS / FID $\downarrow$		PSNR / SSIM $\uparrow$	LPIPS / FID $\downarrow$		PSNR / SSIM $\uparrow$	LPIPS / FID $\downarrow$
Task Specific	NRKNet	26.11 / 0.817	0.223 / 43.96	SwinIR	29.83 / 0.897	0.084 / 8.20	Restormer	39.93 / 0.947	0.198 / 47.24
	DRBNet	25.72 / 0.806	0.182 / 39.37	Restormer	32.71 / 0.960	0.043 / 2.90	Uformer	39.80 / 0.946	0.200 / 47.15
All in One	InsturctIR [56]	23.84 / 0.746	0.329 / 84.88	InsturctIR [56]	31.93 / 0.944	0.061 / 3.77	InsturctIR [56]	35.45 / 0.881	0.356 / 57.45
	DA-CLIP [4]	23.55 / 0.747	0.288 / 67.54	DA-CLIP [4]	30.77 / 0.923	0.079 / 5.58	DA-CLIP [4]	34.18 / 0.838	0.301 / 62.47
	UniRestore [54]	22.91 / 0.724	0.364 / 91.59	UniRestore [54]	30.23 / 0.918	0.080 / 6.40	UniRestore [54]	35.41 / 0.835	0.247 / 56.00
	FoundIR [80]	23.45 / 0.742	0.358 / 89.21	FoundIR [80]	31.43 / 0.930	0.059 / 3.46	FoundIR [80]	37.12 / 0.888	0.266 / 46.53
	DCPT [32]	25.68 / 0.816	0.216 / 42.59	DCPT [32]	31.89 / 0.947	0.050 / 3.08	DCPT [32]	37.07 / 0.881	0.282 / 51.03
	MaskDCPT (Ours)	25.64 / 0.809	0.183 / 38.49	MaskDCPT (Ours)	32.02 / 0.944	0.039 / 2.83	MaskDCPT (Ours)	38.68 / 0.934	0.152 / 29.48
Type	Method	<i>Gaussian Deblur</i>		Method	<i>Demoaic</i>		Method	<i>Demoire</i>	
		PSNR / SSIM $\uparrow$	LPIPS / FID $\downarrow$		PSNR / SSIM $\uparrow$	LPIPS / FID $\downarrow$		PSNR / SSIM $\uparrow$	LPIPS / FID $\downarrow$
Task Specific	SwinIR	32.91 / 0.918	0.077 / 2.34	SwinIR	39.94 / 0.994	0.006 / 1.03	SwinIR	24.89 / 0.888	0.100 / 28.73
	Restormer	33.47 / 0.930	0.064 / 2.21	GRL-S	41.77 / 0.996	0.004 / 0.66	RDNet	26.16 / 0.941	0.091 / 23.64
All in One	InsturctIR [56]	31.37 / 0.884	0.113 / 6.04	InsturctIR [56]	37.08 / 0.977	0.011 / 2.33	InsturctIR [56]	24.69 / 0.843	0.111 / 32.18
	DA-CLIP [4]	30.89 / 0.867	0.128 / 6.45	DA-CLIP [4]	38.12 / 0.990	0.006 / 1.07	DA-CLIP [4]	24.75 / 0.826	0.134 / 38.71
	UniRestore [54]	30.77 / 0.871	0.130 / 6.11	UniRestore [54]	37.99 / 0.990	0.006 / 1.21	UniRestore [54]	24.06 / 0.819	0.155 / 45.28
	FoundIR [80]	32.90 / 0.915	0.073 / 2.59	FoundIR [80]	38.44 / 0.992	0.005 / 0.84	FoundIR [80]	24.71 / 0.876	0.107 / 32.49
	DCPT [32]	32.06 / 0.906	0.087 / 2.81	DCPT [32]	38.11 / 0.991	0.006 / 1.05	DCPT [32]	24.18 / 0.815	0.159 / 31.38
	MaskDCPT (Ours)	33.28 / 0.927	0.057 / 2.13	MaskDCPT (Ours)	38.59 / 0.992	0.005 / 0.72	MaskDCPT (Ours)	25.21 / 0.942	0.095 / 24.41

TABLE II: *12D all-in-one image restoration results* in terms of PSNR $\uparrow$ / SSIM $\uparrow$ / LPIPS $\downarrow$ / FID $\downarrow$. All-in-one network pre-trained with MaskDCPT outperforms task-specific methods in terms of fidelity for deraining, desnowing, raindrop removal, and low-light enhancement. In most restoration tasks, it surpasses task-specific methods in terms of perceptual metrics. All the all-in-one methods are trained on UIR-2.5M to ensure fair comparison.

precision. The performance of the restoration model under 12 degradation is presented in Table II. It can be observed that, (1) compared to abstract CLIP embeddings [4], complex human instructions [56], and the large diffusion model [80, 54], the degradation classification prior to the MaskDCPT-trained model is more effective in addressing the complex all-in-one restoration task. In the context of motion deblurring, the MaskDCPT framework demonstrates an improvement of 1.98 dB, 3.58 dB, and 3.54 dB in PSNR metrics compared to InsturctIR, DA-CLIP, and UniRestore, respectively, while achieving a reduction of 50% in FID metrics. Moreover, MaskDCPT achieves state-of-the-art performance across all other assessed all-in-one tasks. (2) The restoration model trained with MaskDCPT demonstrates superior performance over previous task-specific methods in terms of both fidelity and perceptual quality. For instance, in desnowing task, MaskDCPT surpasses FocalNet by

0.54 dB in PSNR; in low-light enhancement, it exceeds GLARE by 4.78 dB in PSNR. MaskDCPT also exhibits advances over task-specific approaches in perceptual assessments. For the real image denoising task, MaskDCPT achieves a 37.4% reduction in the FID compared to Uformer; and in the raindrop removal task, it obtains a 63.5% reduction in LPIPS relative to UDR-S2Former. (3) The universal restoration method performs similarly to task-specific approaches under global degradation. However, for non-uniform degradation such as haze, motion blur, or defocus blur, task-specific methods perform better.

B. Single-task image restoration

A further analysis is conducted to determine the suitability of MaskDCPT for single-task image restoration pre-training from two perspectives. **i. Zero-shot (ZS):** This evaluates whether MaskDCPT trained models under *12D all-in-one* fine-tuning are

Method	Urban100					Kodak24					BSD68				
	ID			OOD		ID			OOD		ID			OOD	
	$\sigma = 15$	$\sigma = 25$	$\sigma = 50$	$\sigma = 60$	$\sigma = 75$	$\sigma = 15$	$\sigma = 25$	$\sigma = 50$	$\sigma = 60$	$\sigma = 75$	$\sigma = 15$	$\sigma = 25$	$\sigma = 50$	$\sigma = 60$	$\sigma = 75$
AdaIR [57]	34.10	31.68	28.28	26.63	22.60	34.88	32.39	29.22	27.39	23.15	34.01	31.34	28.06	26.47	22.83
DA-RCOT [78]	33.95	31.29	26.36	22.03	16.83	34.73	31.96	26.93	21.82	16.23	33.84	30.91	25.95	21.62	16.29
MocelIR [79]	33.99	31.58	28.21	26.86	23.45	34.85	32.37	29.20	27.72	23.84	33.98	31.34	28.06	26.65	23.26
DFPIR [53]	33.94	31.59	28.29	26.08	21.45	34.77	32.32	29.20	26.66	21.38	33.94	31.29	28.05	25.85	21.28
SwinIR-5D	32.79	30.18	26.52	24.47	19.80	33.89	31.32	27.93	25.36	20.01	33.31	30.59	27.13	24.39	20.11
+ RAM [28]	33.77	31.41	27.95	24.93	20.56	34.51	32.10	28.90	26.03	20.85	33.63	31.06	27.80	24.95	20.82
+ DCPT [32]	33.64	31.14	27.63	24.36	20.04	34.63	32.11	28.86	25.98	20.54	33.82	31.16	27.86	24.49	20.29
+ MaskDCPT (Ours)	33.83	31.39	27.91	25.63	21.31	34.65	32.15	28.93	26.15	21.39	33.78	31.13	27.85	25.51	21.25
NAFNet-5D	33.14	30.64	27.20	25.74	19.93	34.27	31.80	28.62	25.92	18.08	33.67	31.02	27.73	25.90	19.42
+ DCPT [32]	33.64	31.23	27.98	26.30	20.13	34.72	32.28	29.21	27.09	19.99	33.94	31.31	28.12	26.32	19.84
+ MaskDCPT (Ours)	34.11	31.80	28.63	27.03	20.79	34.92	32.49	29.42	27.49	20.31	34.03	31.41	28.21	26.58	20.17
+ MaskDCPT-12D (Ours)	33.86	31.49	28.23	27.28	25.92	34.75	32.31	29.17	28.28	26.89	33.91	31.29	28.04	27.16	25.86
Restormer-5D	33.72	31.26	28.03	25.98	21.89	34.78	32.37	29.08	26.91	23.68	34.03	31.49	28.11	25.31	22.97
+ DCPT [32]	34.14	31.79	28.58	26.31	22.12	34.96	32.49	29.40	27.34	24.08	34.09	31.46	28.25	26.33	23.77
+ MaskDCPT (Ours)	34.17	31.81	28.53	26.94	23.81	34.83	32.36	29.20	27.57	24.30	33.91	31.29	28.06	26.60	23.93
PromptIR-5D	33.27	30.85	27.41	25.74	19.22	34.44	31.95	28.71	26.53	19.41	33.85	31.17	27.89	24.49	19.14
+ RAM [28]	33.64	31.27	27.90	25.94	19.65	34.46	32.01	28.81	26.96	19.80	33.68	31.08	27.81	24.96	19.84
+ DCPT [32]	33.88	31.49	28.15	26.71	22.90	34.78	32.30	29.14	27.58	23.52	33.96	31.32	28.08	25.93	21.77
+ MaskDCPT (Ours)	34.14	31.79	28.52	26.91	23.84	34.81	32.34	29.20	27.56	24.51	33.92	31.30	28.08	26.62	23.91

TABLE III: [ZS] *Gaussian denoising results* in five levels, including in-domain ($\sigma = (15, 25, 50)$) and out-of-domain ($\sigma = (60, 75)$) degradation levels. MaskDCPT improves performance for in-domain (ID) degradation levels. With scaling degradation types and levels in training data, the restoration model can generalize better to out-of-domain (OOD) degradation levels.

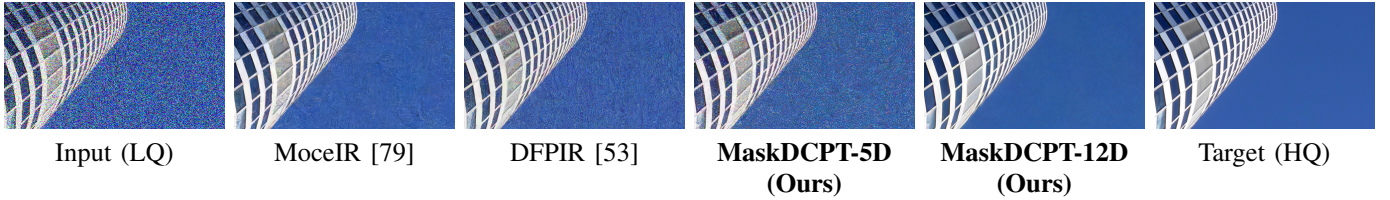


Fig. 6: *Visual comparison on out-of-domain (OOD) scenarios (Gaussian denoising, $\sigma = 75$)*. The MaskDCPT-12D is the only method that effectively removes noise while avoiding the introduction of extraneous artifacts.

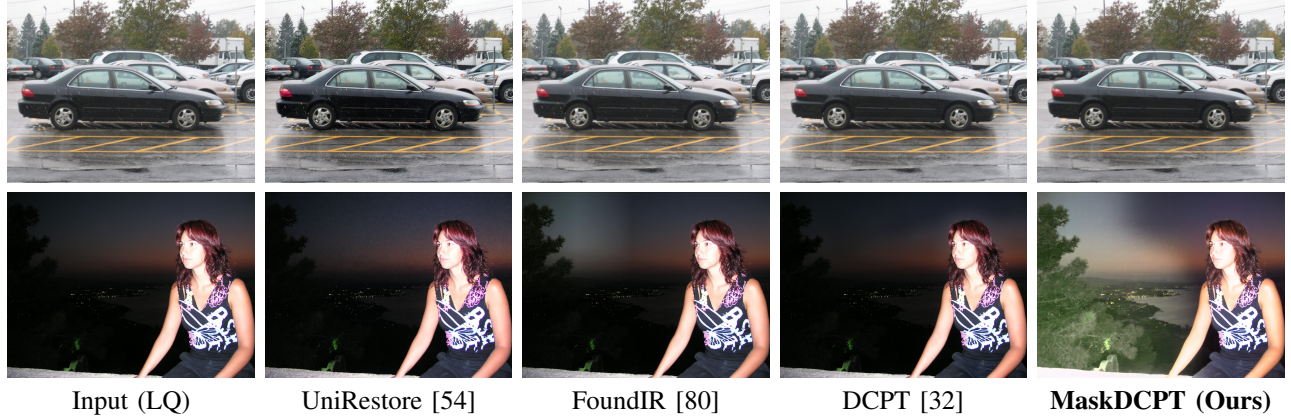
used to solve single tasks without optimization. **ii. Fine-tuning (FT):** This assesses whether the model weights pre-trained with MaskDCPT can be directly used for fine-tuning on single-task image restoration.

[ZS] Implementation details. In zero-shot (ZS) settings, we evaluate the performance of the all-in-one models pre-trained with MaskDCPT on the following: (1) trained tasks at unseen degradation levels, specifically Gaussian denoising in Urban100, Kodak24, and BSD68 datasets. (2) Unseen degradation type within unseen real-world scenarios, including RealBlur-R for motion-deblurring, CUHK and PixelDP for defocus-deblurring, RealRain1K for deraining, Snow100k-real for desnowing, RTTS for dehazing, along with DICM, LIME, MEF, NPE, and VV for low-light enhancement. These real-world datasets have no reference HQ data.

[FT] Implementation details. In fine-tuning (FT) configurations, we train the Restormer [31] model using the Rain13K dataset for image deraining and the GoPro dataset for single image motion deblurring, facilitating a fair comparison with DegAE [26]. The training hyperparameters utilized remain consistent with those employed by Restormer [31]. The key variation lies in the utilization of MaskDCPT pre-trained parameters for model initialization. The fine-tuning process is executed on a single NVIDIA A100 GPU.

[ZS] Unseen degradation levels: Gaussian denoising. Table III and Figure 6 elucidates the Gaussian denoising results of the image restoration model pre-trained with MaskDCPT across various noise levels, including those degradations not encountered during the training phase. (1) The model pre-trained with MaskDCPT evidences substantial improvements across all architectures and testsets, with a particular emphasis on the high-resolution dataset Urban100 [81]. Specifically, MaskDCPT-SwinIR exhibits an enhancement of 1.39 dB over SwinIR in Gaussian denoising with $\sigma = 50$. (2) MaskDCPT displays a marked superiority over existing pre-training methods. Compared to the PromptIR models pre-trained by MaskDCPT and RAM, those pre-trained with MaskDCPT exhibit significant performance enhancements. Notably, within the $\sigma = 50$ and the high-resolution dataset Urban100 [81], MaskDCPT-PromptIR exceeds RAM-PromptIR by 0.97 dB and DCPT-PromptIR by 0.2 dB, respectively. (3) Following exposure to a broader spectrum of degradations, the model demonstrates considerable progress in addressing unseen synthesized levels. In particular, MaskDCPT-NAFNet-12D outperforms MaskDCPT-NAFNet-5D by 5.69 dB in unseen Gaussian noise coefficients, *e.g.*, 75. This performance is attributed to the diverse noise types included within UIR-2.5M, which augment the model's ability to comprehend and mitigate complex noise phenomena.

Degradation	Motion Blur	Defocus Blur	Rain	Snow	Haze	Low-light
Method	PSNR $\uparrow$ / SSIM $\uparrow$ / LPIPS $\downarrow$ / FID $\downarrow$	PIQE $\downarrow$ / BRISQUE $\downarrow$	PSNR $\uparrow$ / SSIM $\uparrow$ / LPIPS $\downarrow$ / FID $\downarrow$	PIQE $\downarrow$ / BRISQUE $\downarrow$	PIQE $\downarrow$ / BRISQUE $\downarrow$	PIQE $\downarrow$ / BRISQUE $\downarrow$
DA-CLIP [4]	12.54 / 0.280 / 0.471 / 89.81	39.48 / 34.98	33.24 / 0.939 / 0.102 / 63.87	31.34 / 24.45	47.67 / 34.90	37.64 / 27.45
InstructIR [56]	34.07 / 0.948 / 0.079 / 24.04	35.31 / 32.21	35.80 / 0.964 / 0.096 / 63.83	33.35 / 24.41	50.97 / 31.45	36.08 / 26.31
UniRestore [54]	30.89 / 0.878 / 0.125 / 42.07	31.15 / 27.53	32.54 / 0.905 / 0.141 / 91.49	32.69 / 27.16	46.88 / 30.95	34.63 / 27.05
FoundIR [80]	29.12 / 0.832 / 0.146 / 43.46	47.40 / 41.74	36.02 / 0.967 / 0.093 / 64.36	33.18 / 26.20	61.14 / 42.26	44.17 / 33.51
DCPT-NAFNet [32]	24.78 / 0.772 / 0.195 / 53.92	38.75 / 37.71	32.82 / 0.914 / 0.159 / 79.27	32.59 / 25.02	52.40 / 37.97	35.48 / 26.97
MaskDCPT-NAFNet (Ours)	32.21 / 0.907 / 0.090 / 22.41	28.61 / 30.19	37.02 / 0.978 / 0.070 / 57.37	30.06 / 23.27	33.98 / 33.21	28.50 / 24.76

TABLE IV: [ZS] *Real-world restoration results* in six real-world degradation types.Fig. 7: *Visual comparison on real-world restoration scenarios.* Zoom in for best view.

Dataset	Method	DeblurGAN	DeblurGANv2	SRN	DMPHN	Restormer	DegAE-Restormer [26]	DCPT-Restormer [32]	MaskDCPT-Restormer (Ours)
GoPro	PSNR $\uparrow$	28.70	29.55	30.26	31.20	32.92	33.03 (+0.11)	33.12 (+0.20)	33.29 (+0.37)
	SSIM $\uparrow$	0.858	0.934	0.934	0.940	0.961	-	0.962	0.964 (+0.03)
HIDE	PSNR $\uparrow$	24.51	26.62	28.36	29.09	31.22	31.43 (+0.21)	31.47 (+0.25)	31.55 (+0.33)
	SSIM $\uparrow$	0.871	0.875	0.915	0.924	0.942	-	0.946 (+0.04)	0.946 (+0.04)
Dataset	Method	SIRR	MSPFN	LPNet	AirNet	Restormer	DegAE-Restormer [26]	DCPT-Restormer [32]	MaskDCPT-Restormer (Ours)
Test100	PSNR $\uparrow$	32.37	33.50	33.61	34.90	36.74	35.39 (-1.35)	37.24 (+0.50)	37.70 (+0.96)
	SSIM $\uparrow$	0.926	0.948	0.958	0.977	0.978	0.972 (-0.06)	0.980 (+0.02)	0.984 (+0.06)

TABLE V: [FT] *Single Image Motion Deblurring results* in the single-task setting on the GoPro dataset. *Image Deraining results* in the single-task setting on the Test100 dataset.

[ZS] Unseen degradation types: real-world blur and weather. Table IV and Figure 7 illustrate the superior generalization capabilities of MaskDCPT in real-world scenarios, significantly outperforming all-in-one restoration methodologies. According to the quantitative metrics, MaskDCPT attained the majority of the state-of-the-art results, notably achieving the lowest Fréchet Inception Distance (FID) in the motion blur task, and delivered superior performance across the other five real-world environments. The visual output delineates that (1) the methods grounded in degradation classification (DCPT [32] and our MaskDCPT) are adept at eliminating small disturbances, *e.g.*, rain and snow from images, unlike the Diffusion-based methods [54, 80]. (2) The integration of mask processing enhances the model’s capacity to discern and ameliorate localized degradations. Although DCPT [32] is effective in globally removing rain and snow, it cannot address local low-light conditions. Our MaskDCPT adeptly resolves this problem by accurately illuminating these regions. More visual comparisons are shown in the supplementary.

[FT] Single-task degradation: motion blur and rain. MaskDCPT is suitable for pre-training on a single task. Table V shows

that Restormer pre-trained with MaskDCPT outperforms 0.37 dB on GoPro. MaskDCPT remains an appropriate approach for pre-training on image deraining tasks. In contrast, DegAE [26] exhibits a reduced performance in image deraining. MaskDCPT exhibits greater universality.

C. Image restoration on mixed degradation

Dataset. We use the UIR-2.5M-mixed as a training dataset, and test our model on CDD and LoL-Blur. The testset comprises prevalent degradation combinations, including low-light, haze, rain-streaks, and snow. We conduct evaluations exclusively on three-mixed degradation to illustrate the advantages of MaskDCPT in restoring intricate degradation mixtures.

Implementation details. We use the NAFNet pre-trained by MaskDCPT on 5D all-in-one restoration datasets to initialize model. We use the AdamW optimizer with the initial learning rate 3×10^{-4} gradually reduced to 1×10^{-6} with the cosine annealing schedule to train our image restoration models. The training runs for 750k iters with batch size 32 on 4 NVIDIA L40 GPUs.

Method	L + H + RS	L + H + S	L + B + N
DACLIP [4]	25.86 / 0.797 / 0.210 / 25.12	25.22 / 0.800 / 0.205 / 28.87	26.45 / 0.862 / 0.147 / 11.48
InstructIR [56]	24.84 / 0.777 / 0.233 / 28.71	24.32 / 0.760 / 0.279 / 40.33	26.33 / 0.860 / 0.163 / 17.31
OneRestore [6]	25.18 / 0.799 / 0.165 / 24.85	25.28 / 0.802 / 0.148 / 24.90	25.02 / 0.788 / 0.227 / 34.32
MoceIR [79]	25.41 / 0.801 / 0.213 / 30.04	25.40 / 0.802 / 0.208 / 28.51	24.28 / 0.753 / 0.317 / 43.28
DCPT-NAFNet [32]	25.76 / 0.817 / 0.188 / 23.01	25.90 / 0.819 / 0.174 / 23.56	25.56 / 0.829 / 0.181 / 23.69
MaskDCPT-NAFNet (Ours)	26.24 / 0.811 / 0.151 / 22.80	26.38 / 0.814 / 0.143 / 22.90	27.13 / 0.881 / 0.131 / 10.27

TABLE VI: *Mixed degraded image restoration results* on CDD [6] and LoL-Blur [82] in terms of PSNR $\uparrow$ / SSIM $\uparrow$ / LPIPS $\downarrow$ / FID $\downarrow$. All methods are trained on UIR-2.5M-mixed to ensure fair comparison.



Fig. 8: *Visual comparison on mixed degradation scenarios*. MaskDCPT can restore the illumination globally.

Mask ratio (%)	Dehazing	Deraining	Denoising	Deblurring	Low-light
0	30.93	37.18	31.27	28.86	23.09
25	31.93	38.84	31.28	29.10	25.71
50	32.71	39.12	31.30	29.99	26.30
75	32.66	39.08	31.24	29.84	26.00

TABLE VII: Ablations of the mask ratio.

Patch size	Dehazing	Deraining	Denoising	Deblurring	Low-light
1	29.33	36.12	31.09	27.89	23.37
4	31.99	39.03	31.28	29.74	26.11
16	32.71	39.12	31.30	29.99	26.30
32	32.17	38.88	31.29	29.68	25.75

TABLE VIII: Ablations of masked patch size.

Results of restoration on mixed degradation are displayed in Table VI. MaskDCPT can deliver substantial performance enhancements to the restoration model in mixed degradation scenarios. Compared with OneRestore [6], MaskDCPT demonstrates a PSNR improvement of 1.06 dB for mixed degradations involving low-light, haze and rain degradation, and 2.11 dB for those involving low-light, blur, and noise degradation. For compelling evidence, Figure 8 provides a visual comparison of image restoration in three composite degradation samples (low-light + haze + rain). NAFNet pre-trained with MaskDCPT can restore more natural result from mixed-degraded image and fully preserve image texture and detail such as lighting and building textures.

D. Ablation studies

Our conference paper [32] has demonstrated the necessity of decoder architecture, multi-scale feature extraction, training stages, and pre-training it self in DCPT through the performance of several ablation experiments. We hereby comprehensively analyze our newly added mask mechanism. The ablations are performed with PromptIR [2] in the 5D all-in-one image restoration task, in terms of PSNR $\uparrow$.

Impact of mask ratio. As shown in Table VII, it was observed that selecting a mask ratio of 50% optimizes restoration performance. In contrast, when the mask ratio is reduced to 0, MaskDCPT reverts to DCPT [32], thus losing its ability to train simultaneously for degradation discrimination and image reconstruction, resulting in a notable performance decline.

Impact of masked patch size. Refer to Table VIII, a patch size of 16 is optimal for MaskDCPT. In instances where the patch size is adjusted to 1, the masking approach aligns with

RAM [28]. Our findings indicate that employing a patch size of 1 during pre-training with degradation classification can notably diminish restoration performance. This occurs because pixel-level masks disrupt the distribution of degradation information throughout the image, thereby impeding the model’s ability to effectively detect degradation, which ultimately impacts the restoration performance.

Masking method	Dehazing	Deraining	Denoising	Deblurring	Low-light
square	32.18	38.94	30.84	29.07	25.82
block-wise	32.22	39.00	30.98	29.27	25.90
random	32.71	39.12	31.30	29.99	26.30

TABLE IX: Ablations of masking methods.

Impact of masking methods. Following SimMIM [19], we conducted ablations involving a variety of masking methods. As evidenced by the results presented in Table IX, the random masking strategy exerts optimal performance in the 5D all-in-one image restoration task. This observation can be attributed to the inherently pixel-intensive nature of image restoration tasks, which necessitate the model’s proficiency in processing various image regions. The application of random masks improves the model’s ability to fit the distribution of pixels between disparate image regions, thus markedly increasing restoration performance.

E. Discussions

Restoration performance as the degradation classification accuracy changes. The results demonstrate a direct correlation between enhancements in the accuracy of degradation classification during pre-training and subsequent improvements in the

Method	MaskDCPT iterations	0	25k	50k	75k	100k
SwinIR [29]	Initial DC Acc. (%)	54	69	82	89	94
	PSNR (dB)	25.04	26.10	27.58	28.44	29.21
NAFNet [30]	Initial DC Acc. (%)	52	90	95	97	98
	PSNR (dB)	27.77	29.40	30.95	31.88	32.09
Restormer [31]	Initial DC Acc. (%)	60	92	97	99	99
	PSNR (dB)	27.60	29.34	30.76	31.63	31.98
PromptIR [2]	Initial DC Acc. (%)	56	90	94	97	98
	PSNR (dB)	28.11	29.77	30.43	31.50	31.88

TABLE X: All-in-one restoration performance improved as the degradation classification accuracy increased. The PSNR are averaged among 5 tasks in 5D all-in-one restoration. “DC” donates the degradation classification.

network’s all-in-one restoration performance. As delineated in Table X, there is a notable improvement in the performance of restoration models, concomitant with an increase in the initial degradation classification accuracy. This correlation implies that the effectiveness of MaskDCPT is largely attributable to its facilitation of degradation classification prior to the initiation of restoration training.

Types	Methods	w/o MaskDCPT	w MaskDCPT
5D	Restormer [31]	27.60 / 0.112	31.98 / 0.055
	+ instructs [56]	30.11 / 0.083	31.93 / 0.056
	+ frequency [57]	30.09 / 0.089	31.99 / 0.055
	+ MoE [79]	30.62 / 0.079	32.07 / 0.050
9D	Restormer [31]	27.14 / 0.139	31.37 / 0.058
	+ instructs [56]	29.67 / 0.094	31.40 / 0.059
	+ frequency [57]	29.20 / 0.118	31.32 / 0.061
	+ MoE [79]	29.46 / 0.101	31.43 / 0.056
12D	Restormer [31]	26.88 / 0.196	30.79 / 0.061
	+ instructs [56]	29.03 / 0.140	30.80 / 0.061
	+ frequency [57]	28.26 / 0.173	30.74 / 0.063
	+ MoE [79]	28.71 / 0.159	30.91 / 0.058

TABLE XI: 5D (H, RS, GN, MB, LL) restoration performance in terms of PSNR $\uparrow$ / LPIPS $\downarrow$ of degradation-aware architectures as influenced by training methods and scaled degradation types.

What do degradation-aware architectures bring to? MaskDCPT has been shown to be highly effective in improving model performance. Furthermore, a baseline model trained with MaskDCPT demonstrates superior performance compared to degradation-aware architectures. We wonder: “What do degradation-aware architectures bring to restoration performance?” We examined variations in the restoration performance of three degradation-aware architectures: instruction [56], frequency [57], and Mixture of Experts (MoE) [79] as influenced by changes in training methods and increased degradation types. Restormer [31] serves as the baseline model. The experimental results are presented in Table XI. **(1)** When training models from scratch, degradation-aware architectures can provide specific performance enhancements. **(2)** However, as the types of degradation increase, the performance of the model experiences varying degrees of decline. Both MaskDCPT and degradation-aware architectures can mitigate this performance decline. From a network architecture perspective, the instruction-based

methodology emerges as the most effective means of alleviating this performance decline. **(3)** After training with MaskDCPT, degradation-aware architectures consistently approximate the baseline performance. It suggests that such designs may not be able to increase the model performance ceiling after fully converged training.

VI. CONCLUSION

This paper first validates that randomly initialized restoration models achieve baseline degradation classification performance while preserving robustness with masked input images. Furthermore, models trained for all-in-one restoration exhibit superior classification accuracy. To enhance this efficacy and robustness, we introduce MaskDCPT and experimentally demonstrate its effectiveness in universal image restoration. By integrating degradation classification priors with image distribution learning, MaskDCPT enhances a 3-4 dB PSNR gain in all-in-one restoration and 35 % less PIQE in real-world scenarios for restoration models. In addition, we gather an extensive dataset for universal image restoration, UIR-2.5M, which is able to improve the generalization of restoration models when addressing unseen degradation. In future work, efforts will be directed to improve the generalization of restoration models in the presence of unseen and complex degradation.

REFERENCES

- [1] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *CVPR*, 2022.
- [2] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *NeurIPS*, 2023.
- [3] Yang Ai, Huaibo Huang, Xiaoqiang Zhou, Jiexiang Wang, and Ran He. Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25432–25444, 2024.
- [4] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for universal image restoration. In *The Twelfth International Conference on Learning Representations*, 2023.
- [5] Dian Zheng, Xiao-Ming Wu, Shuzhou Yang, Jian Zhang, Jian-Fang Hu, and Wei-shi Zheng. Selective hourglass mapping for universal image restoration based on diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [6] Yu Guo, Yuan Gao, Yuxu Lu, Huilin Zhu, Ryan Wen Liu, and Shengfeng He. Onerestore: A universal restoration framework for composite degradation. In *European Conference on Computer Vision*, pages 255–272. Springer, 2025.
- [7] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *TIP*, 2007.

- [8] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.
- [9] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7769–7778, 2020.
- [10] Xiaozhong Ji, Guangpin Tao, Yun Cao, Ying Tai, Tong Lu, Chengjie Wang, Jilin Li, and Feiyue Huang. Frequency consistent adaptation for real world super resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1664–1672, 2021.
- [11] Cong Wang, Jinshan Pan, Wei Wang, Jiangxin Dong, Mengzhu Wang, Yakun Ju, and Junyang Chen. Promptrestorer: A prompting image restoration method with degradation perception. *Advances in Neural Information Processing Systems*, 36, 2024.
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.
- [13] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Nee-lakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Dorsch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.
- [16] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.
- [17] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.
- [18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- [19] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.
- [20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [23] Keyu Tian, Yi Jiang, Chen Lin, Liwei Wang, Zehuan Yuan, et al. Designing bert for convolutional networks: Sparse and hierarchical masked modeling. In *The Eleventh International Conference on Learning Representations*, 2022.
- [24] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020.
- [25] Hongyang Wei, Shuaizheng Liu, Chun Yuan, and Lei Zhang. Perceive, understand and restore: Real-world image super-resolution with autoregressive multimodal generative models. *arXiv preprint arXiv:2503.11073*, 2025.
- [26] Yihao Liu, Jingwen He, Jinjin Gu, Xiangtao Kong, Yu Qiao, and Chao Dong. Dega: A new pretraining paradigm for low-level vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23292–23303, June 2023.
- [27] Haoyu Chen, Jinjin Gu, Yihao Liu, Salma Abdel Magid, Chao Dong, Qiong Wang, Hanspeter Pfister, and Lei Zhu. Masked image training for generalizable deep image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1703, 2023.
- [28] Chu-Jie Qin, Rui-Qi Wu, Zikun Liu, Xin Lin, Chun-Le Guo, Hyun Hee Park, and Chongyi Li. Restore anything with masks: Leveraging mask image modeling for blind all-in-one image restoration. In *European Conference on Computer Vision*, pages 364–380. Springer, 2024.
- [29] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.
- [30] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 17–33, Cham, 2022. Springer Nature Switzerland.
- [31] 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 *CVPR*, 2022.
- [32] JiaKui Hu, Lujia Jin, Zhengjian Yao, and Yanye Lu. Universal image restoration pre-training via degradation classification. *arXiv preprint arXiv:2501.15510*, 2025.
- [33] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–192, 2021.
- [34] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3517–3526, June 2021.
- [35] 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*, pages 5769–5780, 2022.
- [36] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17683–17693, June 2022.
- [37] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22367–22377, June 2023.
- [38] Lujia Jin, Qing Guo, Shi Zhao, Lei Zhu, Qian Chen, Qiushi Ren, and Yanye Lu. One-pot multi-frame denoising. *International Journal of Computer Vision*, 132(2):515–536, 2024.
- [39] JiaKui Hu, Zhengjian Yao, Lujia Jin, Hangzhou He, and Yanye Lu. Enhancing image restoration transformer via adaptive translation equivariance. *arXiv preprint arXiv:2506.18520*, 2025.
- [40] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [42] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.
- [43] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018.
- [44] Xiaoyu Xiang Yawei Li, Yuchen Fan, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.
- [45] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3175–3185, 2020.
- [46] Qingnan Fan, Dongdong Chen, Lu Yuan, Gang Hua, Nenghai Yu, and Baoquan Chen. A general decoupled learning framework for parameterized image operators. *TPAMI*, 2019.
- [47] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer, 2021.
- [48] Jinghao Zhang, Jie Huang, Mingde Yao, Zizheng Yang, Hu Yu, Man Zhou, and Feng Zhao. Ingredient-oriented multi-degradation learning for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5825–5835, 2023.
- [49] 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*, pages 8748–8763. PMLR, 2021.
- [50] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- [51] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.
- [52] Haijin Zeng, Xiangming Wang, Yongyong Chen, Jingyong Su, and Jie Liu. Vision-language gradient descent-driven all-in-one deep unfolding networks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7524–7533, 2025.
- [53] Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, and Chao Ren. Degradation-aware feature perturbation for all-in-one image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28165–28175, 2025.
- [54] I Chen, Wei-Ting Chen, Yu-Wei Liu, Yuan-Chun Chiang, Sy-Yen Kuo, Ming-Hsuan Yang, et al. Unirestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17969–17979, 2025.
- [55] Mo Zhou, Keren Ye, Mauricio Delbracio, Peyman Milanfar, Vishal M Patel, and Hossein Talebi. Unires: Universal image restoration for complex degradations. *arXiv preprint arXiv:2506.05599*, 2025.
- [56] Radu Timofte Marcos V. Conde, Gregor Geigle. High-quality image restoration following human instructions, 2024.
- [57] Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Shahbaz Khan. Adair: Adaptive all-in-one image restoration via frequency mining and modulation. *arXiv preprint arXiv:2403.14614*, 2024.
- [58] Xu Zhang, Jiaqi Ma, Guoli Wang, Qian Zhang, Huan

- Zhang, and Lefei Zhang. Perceive-ir: Learning to perceive degradation better for all-in-one image restoration. *IEEE Transactions on Image Processing*, 2025.
- [59] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.
- [60] Peng Gao, Teli Ma, Hongsheng Li, Ziyi Lin, Jifeng Dai, and Yu Qiao. Mcmae: Masked convolution meets masked autoencoders. *Advances in Neural Information Processing Systems*, 35:35632–35644, 2022.
- [61] Man Yao, Xuerui Qiu, Tianxiang Hu, Jiakui Hu, Yuhong Chou, Keyu Tian, Jianxing Liao, Luziwei Leng, Bo Xu, and Guoqi Li. Scaling spike-driven transformer with efficient spike firing approximation training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [62] Qi Han, Yuxuan Cai, and Xiangyu Zhang. Revcolv2: Exploring disentangled representations in masked image modeling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [63] Wenbo Li, Xin Lu, Shengju Qian, and Jiangbo Lu. On efficient transformer-based image pre-training for low-level vision. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 1089–1097. International Joint Conferences on Artificial Intelligence Organization, 8 2023. Main Track.
- [64] 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. *TPAMI*, 2019.
- [65] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *TIP*, 2018.
- [66] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, 2018.
- [67] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017.
- [68] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018.
- [69] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021.
- [70] Yuxuan Cai, Yizhuang Zhou, Qi Han, Jianjian Sun, Xiangwen Kong, Jun Li, and Xiangyu Zhang. Reversible column networks. In *International Conference on Learning Representations*, 2023.
- [71] 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.
- [72] Keyu Tian, Yi Jiang, Qishuai Diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing bert for convolutional networks: Sparse and hierarchical masked modeling. *arXiv preprint arXiv:2301.03580*, 2023.
- [73] Benjamin Graham and Laurens Van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017.
- [74] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhong Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1787, 2023.
- [75] 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*, pages 1357–1366, 2017.
- [76] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [77] Xiaole Tang, Xin Hu, Xiang Gu, and Jian Sun. Residual-conditioned optimal transport: towards structure-preserving unpaired and paired image restoration. In *International Conference on Machine Learning*, pages 47757–47777. PMLR, 2024.
- [78] Xiaole Tang, Xiang Gu, Xiaoyi He, Xin Hu, and Jian Sun. Degradation-aware residual-conditioned optimal transport for unified image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [79] Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12753–12763, 2025.
- [80] Hao Li, Xiang Chen, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Foundir: Unleashing million-scale training data to advance foundation models for image restoration. *arXiv preprint arXiv:2412.01427*, 2024.
- [81] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015.
- [82] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *European conference on computer vision*, pages 573–589. Springer, 2022.