DegAE: A New Pretraining Paradigm for Low-level Vision

Yihao Liu[1,2,3] Jingwen He[1] Jinjin Gu[1,4] Xiangtao Kong[1,2,3] Yu Qiao[1,2] Chao Dong[2,1*]

[1] Shanghai Artificial Intelligence Laboratory [2] ShenZhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences [3]University of Chinese Academy of Sciences [4]The University of Sydney

Abstract

Self-supervised pretraining has achieved remarkable success in high-level vision, but its application in low-level vision remains ambiguous and not well-established. What is the primitive intention of pretraining? What is the core problem of pretraining in low-level vision? In this paper, we aim to answer these essential questions and establish a new pretraining scheme for low-level vision. Specifically, we examine previous pretraining methods in both high-level and low-level vision, and categorize current low-level vision tasks into two groups based on the difficulty of data acquisition: low-cost and high-cost tasks. Existing literature has mainly focused on pretraining for low-cost tasks, where the observed performance improvement is often limited. However, we argue that pretraining is more significant for high-cost tasks, where data acquisition is more challenging. To learn a general low-level vision representation that can improve the performance of various tasks, we propose a new pretraining paradigm called degradation autoencoder (DegAE). DegAE follows the philosophy of designing pretext task for self-supervised pretraining and is elaborately tailored to low-level vision. With DegAE pretraining, SwinIR achieves a 6.88dB performance gain on image dehaze task, while Uformer obtains 3.22dB and 0.54dB improvement on dehaze and derain tasks, respectively.

1. Introduction

With the phenomenal success of self-supervised pretraining in natural language processing (NLP), a large number of attempts have also been proposed in the field of computer vision [20,21,66,67]. The idea behind self-supervised pretraining is to learn a general visual representation by devising an appropriate pretext task that does not rely on any manual annotation. Owing to large-scale pretraining, models with a voracious appetite for data can alleviate the overfitting problem and achieve further improvement.

* Corresponding author. Email: chao.dong@siat.ac.cn.

Recently, referring to the philosophy of masked language modeling (MLM) in NLP [27,51], masked image modeling (MIM) [20,67] has been proposed and proven to be extraordinarily effective in high-level vision tasks, *e.g*., image classification, object detection, and image segmentation. However, the notion of low-level vision pretraining is not yet well-established, due to the distinctions between high-level and low-level vision tasks. Specifically, the representative high-level vision tasks take fixed-size images as inputs and predict manually annotated labels as targets [15, 23], while most low-level vision methods accept low-quality (LQ) images as inputs and produce high-quality (HQ) images as targets [31, 78]. More importantly, the annotation manner in low-level vision is quite different. To obtain LQ-HQ pairs, a wide range of tasks choose to synthesize input LQ images from collected HQ images, such as classical super-resolution [11] and Gaussian denoise [77]. Based on the difficulty of paired-data acquisition, we can roughly categorize low-level vision tasks into two groups: 1) *low-cost task*: tasks with low-cost data acquisition (*e.g*., super-resolution), and 2) *high-cost task*: tasks with high-cost data acquisition (*e.g*., dehaze). This analysis is absent in existing low-level vision literatures [4, 7, 34]. They only consider low-cost tasks and simply adopt a straightforward pretraining strategy that has the same objectives as the downstream tasks. Such a pretraining paradigm lacks generality and only brings marginal improvement. In this paper, we claim that pretraining could potentially be more effective for high-cost tasks and that a new pretraining paradigm tailored to low-level vision would be highly beneficial.

To this end, we devise a novel pretraining paradigm for low-level vision. Since the goal of low-level vision is to process LQ images with various degradations, we propose a degradation autoencoder (DegAE) to achieve content-degradation disentanglement and generation. DegAE accepts an input image with degradation $\mathcal{D}_1$ and a reference image with degradation $\mathcal{D}_2$. It attempts to transfer the degradation $\mathcal{D}_2$ of the reference image to the input image, obtaining an output image with input image content, but with degradation $\mathcal{D}_2$, as described in Fig. 1. Through such a learning paradigm, the model is expected to learn both

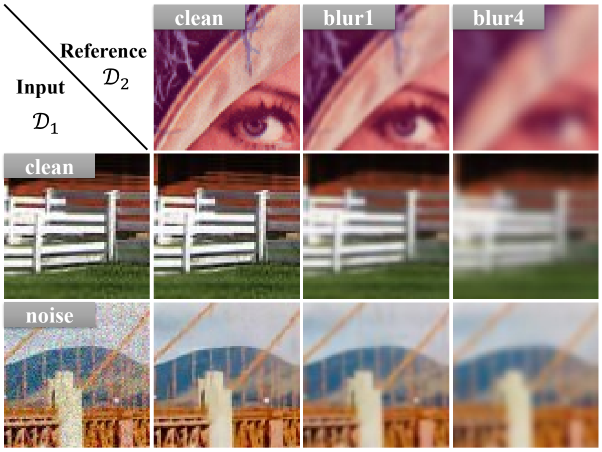

Figure 1. Example results of DegAE pretraining. For instance, given an input noise image and a reference blur image, DegAE attempts to transfer the blur degradation to the input image. More visual examples are illustrated in the supplementary file.

natural image representations and degradation information, which are the key components in low-level vision. Our approach follows the philosophy of designing pretext task for self-supervised pretraining [20, 27]. Firstly, the pretext task should not depend on the downstream tasks, in order to achieve the generality and transferability of the pretrained representations. Secondly, the pretext task should be carefully designed to exploit internal structures of data.

To validate the effectiveness of DegAE pretraining, we choose three representative backbone models (SwinIR [38], Uformer [64] and Restormer [71]) to conduct experiments. The results suggest that DegAE pretraining can significantly improve the model performance. For example, SwinIR yields a 6.88dB gain on image dehaze task (SOTS) and a 1.27dB gain on image derain task (Rain100L). Uformer obtains 3.22dB and 0.54dB improvement on image dehaze and derain task (Test100). Restormer achieves 0.43dB performance improvement on image motion deblur task (GoPro), respectively. As expected, we also observe incremental improvement on low-cost tasks – SR and denoise tasks. We believe our efforts can help to bridge the gap between high-level and low-level vision tasks and improve the performance of various low-level vision tasks.

2. Related Work

Image Restoration. The rise of deep learning has led to significant developments in image restoration [2, 11, 29, 77]. The purpose of image restoration is to reconstruct high-quality natural images from observed corrupted images. Typical image restoration tasks include image deblur, denoise, dehaze, derain, super-resolution, *etc*. [2, 11, 14, 29, 30, 70, 77, 78]. A pioneer work SRCNN [11] first introduced convolutional neural networks (CNN) to perform super-resolution. Zhang *et al*. [77] proposed the first deep denoise method DnCNN. DehazeNet [2] and MSCNN [54] were the forerunners in applying the deep learning-based method to image dehaze. For image motion deblur, DeblurGAN [29] and DeblurGAN-v2 [30] leveraged generative adversarial learning to achieve more realistic results. Afterwards, more advanced methods like SRN [60], SPAIR [49] and NAFNet [5] were proposed. For image derain, the representative methods include DerainNet [16], PreNet [53] and MSPFN [25]. A multitude of follow-up works have been proposed in low-level vision tasks and achieved continuous improvements.

Vision Transformer. Transformer [62] has dominated the model design in natural language processing (NLP). Due to its powerful representation learning capabilities, many attempts have been made to explore Transformer in various vision tasks, such as image recognition [15], segmentation [59] and object detection [3]. Along with high-level vision tasks, Transformer-based methods are also deployed in low-level vision tasks. Based on ViT, Chen *et al*. [4] proposed IPT for image restoration. Liang *et al*. [38] presented a stronger baseline model SwinIR based on Swin Transformers [44]. To facilitate long-range pixel dependencies and multi-scale local-global representation learning, Uformer [64] and Restormer [71] were proposed. They both adopted an encoder-decoder design to achieve higher efficiency, establishing new state-of-the-art baselines on various image restoration tasks.

Self-supervised Pretraining. In the field of NLP, with the power of Transformer and billion-scale data, self-supervised pretraining has become a default option. The tacit recipe is to pretrain on a large corpus and then fine-tune on a smaller task-specific dataset. Masked language modeling and its variants have been proven successful for pretraining, *e.g*., GPTs [51, 52] and BERT [27]. As for computer vision, diverse pretext tasks also have been invented to learn visual representation, *e.g*., jigsaw puzzle solving [48], rotation prediction [18], instance discrimination [21, 66]. Recently, masked image modeling (MIM) has been proposed for vision tasks, where the pioneer works are MAE [20] and SimMIM [67]. Experiments show that MIM pretraining can learn abstract and discriminative representations, achieving promising transfer learning results. As for low-level vision tasks, a few pretraining methods have been proposed [4, 7, 34, 41]. For instance, IPT [4] proposed multi-task restoration pretraining and EDT [34] proposed multi-related-task restoration pretraining. However, the motivation and paradigm of these pretraining methods are ambiguous, compared to the prevalent pretraining in high-level vision. In this paper, we propose a novel pretraining paradigm tailored to low-level vision – degradation autoencoder (DegAE), which is more general for various downstream tasks.

3. Rethinking Pretraining in Computer Vision

3.1. Revisiting High-level Vision Pretraining

The Primitive Intention of Pretraining. In high-level vision tasks, manual labeling is expensive (*e.g*., objective detection and image segmentation), resulting in limited labeled data for model training. As deep learning-based architectures are becoming more powerful and data-hungry, they can easily overfit to limited training data, even to hundreds of millions data [9, 15]. To address this issue, pretraining on large-scale datasets (*e.g*., ImageNet) is adopted [10,18,20,21,48,66,67,79]. It aims to learn an effective and general visual representation that can be transferred to various downstream tasks, thus alleviating the overfitting problem. SimMIM [67] and MAE [20] both present masked image modeling (MIM) for visual representation learning. These self-supervised pretraining methods have proven to be scalable and shown significant effect on diverse well-known benchmark datasets [9,26,39,80]. Overall, pretraining has proven to be a powerful tool for learning visual representations in scenarios where labeled data is scarce. By designing a pretext task, a transferable representation can be learned to complement the downstream finetuning.

Can We Directly Borrow Masked Image Modeling for Low-Level Vision? Low-level vision tasks require more continuous and spatial information at the pixel-level, whereas high-level vision tasks are concerned with discrete and abstract semantic information. However, the pretraining method MAE [20] is not suitable for low-level vision tasks due to its pretext task design and backbone architecture. MAE masks random patches up to a masking ratio of 75% and reconstructs the missing patches, which results in a significant loss of high-frequency information, such as edges, textures, and structures. Furthermore, MAE is designed based on ViT [15], which directly splits the input image into 16×16 patches and transforms them into a sequence of linear embeddings. The aggressive masking strategy and rough patch-splitting of MAE lead to severe artifacts and over-smoothed results. To address this limitation, current Transformer-based low-level models [38,64,71] still adopt CNN for pre/post-processing. To investigate the applicability of pretraining methods designed for high-level vision tasks, we finetuned a ViT-based autoencoder initialized from MAE on the image dehaze task and also trained the autoencoder from scratch for comparison. Despite some improvement over training from scratch (achieving 26.08dB and 26.12dB on SOTS), the results were still far below the state-of-the-art results (36.39dB in FFA-Net and 37.84dB in DehazeFormer). Furthermore, the visual results, as shown in the supplementary file, were over-smoothed and contained artifacts. This experiment highlights the limitations of directly applying high-level vision pretraining methods to low-level vision.

3.2. Rethinking Low-level Vision Pretraining

Analysis on Low-level Vision tasks. Before we examine the current low-level pretraining methods, let us pay attention to some important characteristics of low-level vision tasks. According to the paired-data acquisition process, we can roughly classify low-level vision tasks into two categories: 1) *low-cost tasks*: tasks with low-cost data acquisition; 2) *high-cost tasks*: tasks with high-cost data acquisition. For the first group, the paired training data can be easily synthesized by simple and cheap predefined operations. For instance, image super-resolution (SR) task can be accomplished by downsampling high-resolution images using bicubic interpolation, while Gaussian denoising task can be achieved by adding Gaussian noise to clean images. These degradation processes are relatively simple and can be implemented on-the-fly during training with low cost. For the second group, the data acquisition process is relatively expensive. For example, to simulate hazy images, depth information estimation is required, which cannot be naively implemented online. Therefore, hazy-clean image pairs need to be carefully prepared in advance. Our observations suggest that pretraining can provide significant gains for *high-cost tasks*, but only marginal improvements for *low-cost tasks*. Unfortunately, existing pretraining schemes have not taken into account these characteristics of low-level vision tasks, and their design motivations remain unclear, thus limiting their effectiveness in exploiting the full potential of low-level vision pretraining.

Rethinking Low-level Vision Pretraining. Now let us have a closer look at current low-level vision pretraining methods. We provide a summary of prevalent high-level and low-level pretraining methods in Tab. 1. Among the recently proposed low-level vision pretraining methods, IPT [4], EDT [34], and HAT [7] only consider low-cost tasks, such as image super-resolution, Gaussian denoise, and simple derain [1]. Specifically, IPT [4] adopts a multi-task restoration (SR+denoise+derain) pretraining on ImageNet dataset and then performs finetuning for each specific task separately. However, the actual performance gains from pretraining have not been justified. HAT [7] utilizes a single-task restoration pretraining, and finds that pretraining on the ImageNet dataset for $\times4$ SR brings slight improvement (around 0.1dB). EDT [34] proposes a multi-related-task pretraining method that handles several highly related tasks, such as $\times2$, $\times3$, $\times4$ SR, on a partial ImageNet (200k) dataset. Each sub-task (*e.g*., $\times4$ SR) is finetuned on a smaller dataset (*e.g*., DF2K [1, 61]). However, only marginal improvement is observed on the Gaussian denoise task (less than 0.1dB). In summary, these low-level pretraining methods do not achieve significant performance gains on downstream tasks while requiring a substantial amount

[1] Rain model is single and fixed.

Table 1. A summary of high- and low-level pretraining methods. Our DegAE shows similar properties as high-level pretraining methods: the difficulty of data acquisition of downstream tasks is high-cost; the objective of the pretext task is different from the downstream tasks.

		Pretraining		**Finetuning**		
	Method	**Pretext Task**	**The Difficulty of Data Acquisition**	**Downstream Tasks**	**The Difficulty of Data Acquisition**	**Same objective as pretraining?**
High level	MoCo [21] MAE [20] SimMIM [67]	Instance Discrimination Masked image modeling	Low-cost	Image Classification Object detection Segmentation	High-cost	✗
Low level	IPT [4]	Multi-task restoration (SR&Denoise&Derain)	Low-cost	SR Denoise Derain	Low-cost	✔
	EDT [34]	Multi-related-task restoration (SR/Denoise/Derain)	Low-cost	SR Denoise Derain	Low-cost	✔
	HAT [7]	Single-task restoration (SR)	Low-cost	SR	Low-cost	✔
	DegAE (Ours)	Degradation image modeling	Low-cost	Motion Deblur Derain Dehaze	High-cost	✗

of computational resources.

After analyzing current low-level vision pretraining-finetuning paradigms, we have identified two main reasons why they are less significant. Firstly, these paradigms all focus on *low-cost tasks* for downstream finetuning, where training image pairs can be easily created with no limitation. As a result, performance can be improved by simply collecting more clean/high-resolution images or scaling up model size [38]. Therefore, two-stage pretraining-finetuning on the same or different datasets appears redundant. More importantly, low-cost tasks do not typically suffer from severe overfitting problems, rendering pretraining unnecessary. Secondly, the pretraining and downstream finetuning objectives are the same, implying that the learned representations can only benefit tasks involved in pretraining. For a new downstream task, a corresponding new pretraining must be conducted. Therefore, the application scope of these task-specific pretraining methods is very limited.

Summary. In low-level vision tasks, we need to pay more attention to *high-cost tasks*, as these tasks are more prone to overfitting due to the expense of data acquisition. It is crucial to design a pretext task that enables effective representation learning specifically tailored to low-level vision tasks. The pretext task should not be dependent on the downstream tasks, but rather aim to learn a general representation that is beneficial for various downstream tasks.

In this paper, we select three high-cost tasks including dehaze, motion deblur, and complex derain to conduct experiments. Specifically, dehaze requires depth estimation [14, 32, 37]; motion deblur relies on video acquisition and non-trivial blurring operations [29, 33, 47, 76]; complex derain considers the mixture of various rain synthetic models [24, 36, 40, 46, 68, 70], such as additive composite model [36], screen blend model [46], rain model with occlusion [40], depth-aware rain model [24], *etc*. These fixed training image pairs are produced in advance and directly used in our downstream finetuning. Instead of achieving a semantic-level understanding of images by predicting the largely masked information in MAE, we devise a new pretraining paradigm for low-level vision – degradation autoencoder (DegAE). DegAE corrupts the images and then performs implicit reconstruction and generation. This process requires an understanding of natural image representation and degradation information, which are crucial for general low-level vision tasks.

4. DegAE: A New Pretraining Paradigm in Low-level Vision

In this section, we introduce an effective degradation autoencoder (DegAE) for low-level vision representation learning. The schematic illustration is depicted in Fig. 2. We first corrupt a clean image using a sequence of degradation operations. DegAE accepts the corrupted image $I^{\mathcal{D}_1}$ with degradation $\mathcal{D}_1$ and a reference image $I_{ref}^{\mathcal{D}_2}$ with degradation $\mathcal{D}_2$. It aims to transfer the degradation $\mathcal{D}_2$ to the input image, for obtaining an output image $\hat{I}^{\mathcal{D}_2}$ with input image content, but with reference degradation $\mathcal{D}_2$. DegAE has a Transformer-based encoder that operates directly on the degraded input, and a CNN-based decoder that regenerates the transferred output image based on the encoded feature representations. This self-supervised learning paradigm can effectively extract informative representations that contain natural image statistics and degradation information.

Degradation Input. In DegAE, we apply a sequence of degradations on clean images. Generally, the clean image I is first convolved with blur kernel $\mathbf{k}$. After that, noise $\mathbf{n}$ is added. Then JPEG compression with quality q is applied. Specifically, we have $\widetilde{I} = [I \otimes \mathbf{k} + \mathbf{n}]_{JPEG_q}$, where $\widetilde{I}$ is the degraded image. Following [63, 75], in terms of the choices of blur kernel $\mathbf{k}$, we mainly consider isotropic and anisotropic Gaussian filters. For noise $\mathbf{n}$, we adopt addi-

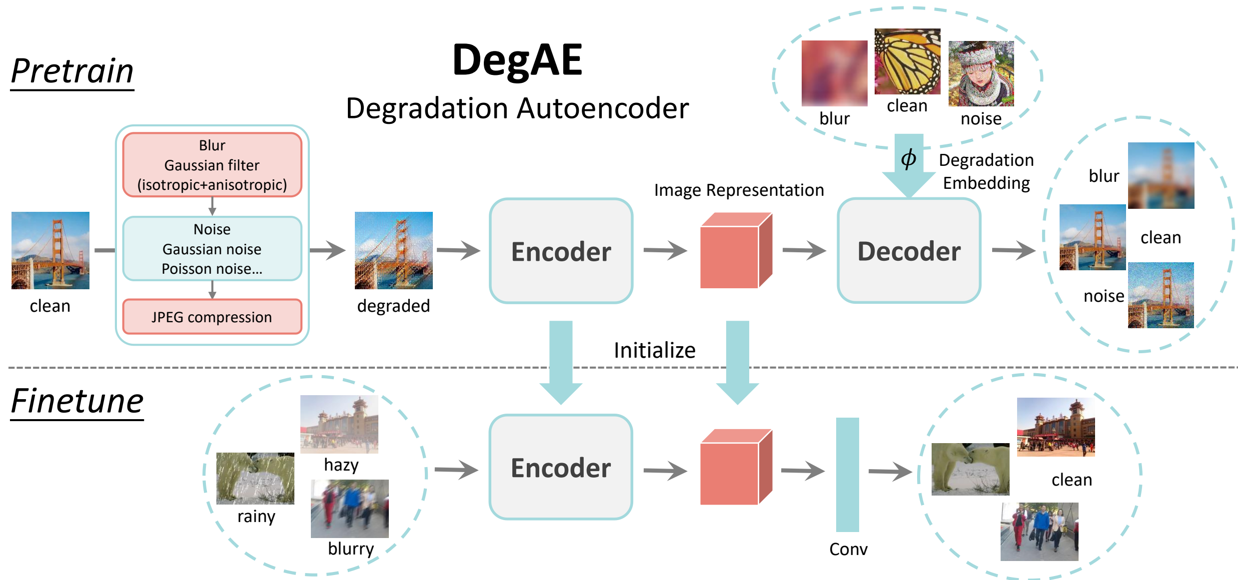

Figure 2. DegAE (Degradation Autoencoder): a new pretraining paradigm for low-level vision. For pretraining, the encoder accepts a degraded input image and outputs the image representation. The degraded input image is synthesized online through a series of degradation operations. The decoder accepts a reference degradation embedding, which is obtained by a degradation representor ϕ. Then, the decoder attempts to transfer the reference degradation to the corrupted input image. During Finetuning, the decoder is replaced by one convolution layer. We finetune the whole network on downstream tasks such as image dehaze, derain and motion deblur.

tive Gaussian noise, Poisson noise, and Speckle noise. Regarding the JPEG compression, we use the PyTorch version: `DiffJPEG` [45]. The degradation level of each degradation type is sampled randomly within a predefined range, which is described in the supplementary file.

Encoder. For any given degraded image, our encoder E produces the low-level feature representation, which will be used to generate diverse outputs in the decoder. At present, there is no unified architecture that can achieve the best results on all low-level vision tasks. Therefore, we use three state-of-the-art Transformer architectures in low-level vision – SwinIR [38], Uformer [64], and Restormer [71] as our encoder. These three architectures have different preferences in handling various tasks. SwinIR mainly performs well on super-resolution and denoise. While Restormer obtains the best performance in derain and dehaze. Uformer could achieve state-of-the-art results in motion deblur. We modify the channel number of the last convolution layer from 3 to 64 for adaptation to the subsequent decoder.

Decoder. Our decoder D accepts the latent feature representation and produces one or more forms of the original clean images. The decoder is a pure CNN architecture that contains four residual blocks [23]. A degradation injection module (implemented referring to [19]) is introduced for the decoder to generate diverse output images. Specifically, the degradation injection module accepts a degradation embedding and then outputs the modulators to modulate the intermediate features of the decoder. Inspired by the analysis of deep representations of SR networks [43], we use a degradation representor ϕ that contains a pretrained SRGAN [31] model and several downsampling layers to produce the degradation embeddings based on the given degraded reference images. Formally, given the degraded input image $I^{\mathcal{D}_1}$ and the degraded reference image $I^{\mathcal{D}_2}_{ref}$, we have $\hat{I}^{\mathcal{D}_2} = D(E(I^{\mathcal{D}_1}), \phi(I^{\mathcal{D}_2}_{ref}))$, where $\hat{I}^{\mathcal{D}_2}$ is the output image, which is expected to be close to the target image $I^{\mathcal{D}_2}$. Note that the reference $I^{\mathcal{D}_2}_{ref}$ could also be a clean image and then the corresponding output image is expected to be clean. In particular, if we set all reference images to clean images, our method will degenerate to previous multi-task restoration pretraining [4], which is a special case of ours. The DegAE decoder is only used in the pretraining stage. It will be replaced by a single convolution layer as the output head during downstream finetuning. The decoder design plays a key role in determining the effectiveness of image representation. The designing philosophy of DegAE is illustrated in the supplementary file.

Reconstruction Target. We adopt four losses to train DegAE: content reconstruction loss $\mathcal{L}_{content}$, perceptual loss $\mathcal{L}_{per}$, adversarial loss $\mathcal{L}_{adv}$, and embedding loss $\mathcal{L}_{embed}$.

Content Reconstruction Loss: For content consistency, we apply a simple Gaussian blur kernel $\mathbf{k}$ on the output images as well as the target images, and then calculate L2 loss between the blurred output image and blurred target image in the pixel space: $\mathcal{L}_{content} = ||I^{\mathcal{D}_2} \otimes \mathbf{k}, \hat{I}^{\mathcal{D}_2} \otimes \mathbf{k}||^2$.

Adversarial Loss: We use generative adversarial learning

Figure 3. Visual results of three low-level vision tasks. We choose three representative backbones (SwinIR, Uformer and Restormer) to verify the effectiveness of DegAE pretraining, since different architectures have their preferences in handling different tasks.

to close the gap between the output distribution and the target distribution. Practically, we adopt the discriminator $\mathbf{D}$ of PatchGAN [81] for adversarial training. $\mathcal{L}_{adv} = [\log \mathbf{D}(I^{\mathcal{D}_2}) + \log(1 - \mathbf{D}(\hat{I}^{\mathcal{D}_2}))]$.

Perceptual Loss: We use VGG19 [57] as the feature extractor, and calculate L2 loss in the feature space: $\mathcal{L}_{per} = ||VGG(I^{\mathcal{D}_2}), VGG(\hat{I}^{\mathcal{D}_2})||^2$.

Embedding Loss: To guarantee that the output image shares similar degradation embedding with the reference degradation, we calculate the L2 loss between their corresponding embedding vectors. The embedding loss is illustrated as $\mathcal{L}_{embed} = ||\phi(I^{\mathcal{D}_2}), \phi(\hat{I}^{\mathcal{D}_2})||^2$.

Finally, the above losses are combined together: $\mathcal{L}_{DegAE} = \lambda_{content} * \mathcal{L}_{content} + \mathcal{L}_{embed} + \mathcal{L}_{per} + \lambda_{adv} * \mathcal{L}_{adv}$, where $\lambda_{content} = 0.1$ and $\lambda_{adv} = 0.005$. These losses are commonly-used in existing GAN-based SR methods. In the supplementary file, we show that GAN loss and perceptual loss are essential for learning complex degradations and $L_{content}$ can maintain the image contents.

5. Experiments

We evaluate the proposed pretraining method on several downstream tasks, including image dehaze, motion deblur, derain, denoise, and super-resolution (SR). In practice, pretraining can bring a large improvement for high-cost tasks (dehaze, motion deblur, and complex derain), but obtains marginal improvement for low-cost tasks (denoise and SR).

Table 2. Quantitative comparisons on dehaze dataset. DegAE pretraining can significantly improve the model performance.

Method	SOTS-ITS	
	PSNR (dB)	SSIM
DCP [22]	16.62	0.818
GFN [55]	22.30	0.880
PFDN [13]	32.68	0.976
GridDehazeNet [42]	32.16	0.984
MSBDN [12]	33.67	0.985
FFA-Net [50]	36.39	0.989
AECR-Net [65]	37.17	0.990
DehazeFormer-B [58]	37.84	0.994
DehazeFormer-M [58]	38.46	0.994
SwinIR	29.83	0.973
DegAE (SwinIR)	**36.71 (+6.88)**	**0.991**
Uformer	31.98	0.984
DegAE (Uformer)	**35.20 (+3.22)**	**0.989**
Restormer	39.01	0.995
DegAE (Restormer)	**39.39 (+0.38)**	**0.995**

This observation is consistent with the analysis conducted in Section 3.2. Due to the space limit, the results of low-cost tasks are described in the supplementary file.

Implementation Details. For pretraining, the learning rate is initialized as 2e-4 and is halved at [50K, 100K, 200K, 300K] iteration. Adam optimizer [28] with β_1 =0.9 and β_2 =0.99 is adopted. We randomly crop 128×128 image patches from DF2K [1, 61] dataset for training. The batch size is set to 2. A total of 600K iterations are executed. After pretraining, we finetune the model on specific downstream datasets. For fairness and convenience, we adopt the same training policy for different backbones. Therefore, the results may observe slight deviations from their original papers, but it does not affect the validation of our method. One can easily exploit more tailored settings for better performance. More details are in the supplementary file.

5.1. Experiment on Image Dehaze

Following [42, 50, 58], the Indoor Training Set (ITS) of RESIDE dataset [32] is adopted for training, which contains a total of 13,990 pairs. The corresponding testing set (SOTS-indoor) consists of 500 indoor images. We compare the quantitative performance of the proposed DegAE pretraining scheme and baselines. Besides, we also report the results of other state-of-the-art methods, including DCP [22], GFN [55], PFDN [13], GridDehazeNet [42], MSBDN [12], FFA-Net [50], AECR-Net [65] and DehazeFormer [58]. Visual results are shown in Fig. 3.

The quantitative results are summarized in Tab. 2. Compared to training from scratch, DegAE pretraining significantly improves the model's dehaze performance, especially for SwinIR and Uformer. The PSNR values of SwinIR and Uformer improve from 29.83dB to 36.71dB and from 31.98dB to 35.20dB, with a performance gain of 6.88dB and 3.22dB, respectively. The results clearly demonstrate the effectiveness of the proposed self-supervised pretraining paradigm. Qualitatively, as shown in Fig. 4, DegAE pretraining can help suppress the generated artifacts, *e.g.*, inhomogeneous background, abnormal colors, and box artifacts. This is due to the fact that the designed pretraining paradigm can enable the model to obtain effective prior visual representations of natural images, making the results closer to the natural clean images. Both quantitative and qualitative results demonstrate the potential of DegAE pretraining.

5.2. Experiment on Image Derain

We train the models on Rain13K dataset, which is newly-adopted in [6, 64, 71, 72]. Rain13K includes 13,712 clean-rain image pairs collected from multiple datasets [16, 35, 36, 46, 70]. We evaluate the derain performance on Rain100L [69], Rain100H [69], Test100 [74], Test1200 [73] and Test2800 [17] datasets. Similar to previous literatures, we calculate the PSNR and SSIM values on the Y channel in the YCbCr color space. We report the results of DegAE along with existing derain methods DerainNet [16], RESCAN [35], PreNet [53], MSPFN [25] and MPRNet [72]. A visual result is shown in the second row of Fig. 3.

From Tab. 3, we can see that DegAE pretraining helps improve the model performance on all five benchmark

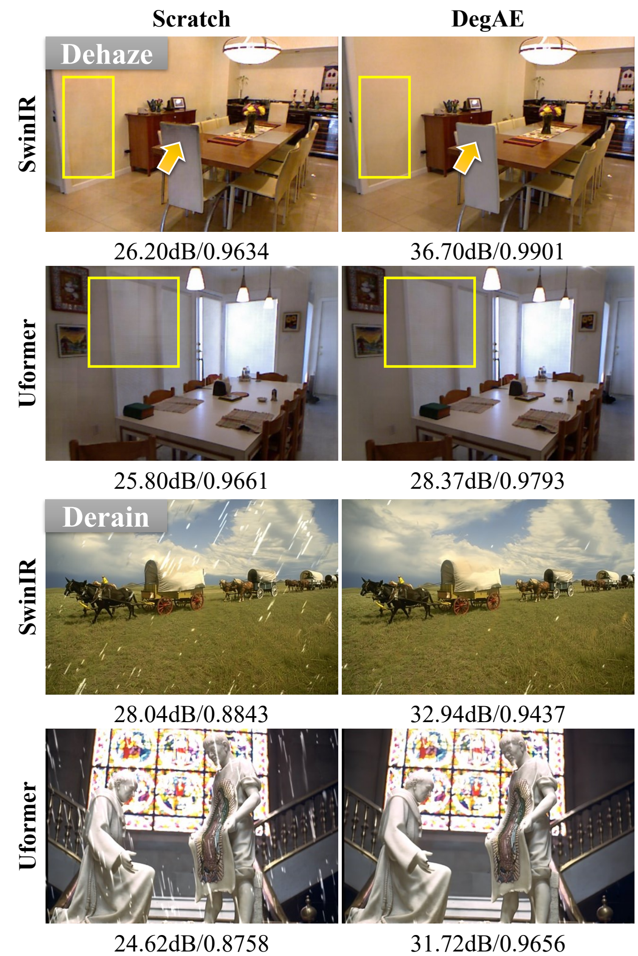

Figure 4. Visual comparison of training from scratch and DegAE pretraining on dehaze and derain effects. DegAE pretraining can reduce the generated artifacts and help remove the haze/rain more thoroughly, compared to training from scratch.

datasets. Specifically, SwinIR yields a 1.27dB improvement on Rain100L dataset with DegAE pretraining. Uformer achieves 0.54dB gain on Test100 dataset with DegAE pretraining. Although Restormer trained from scratch has already achieved state-of-the-art performance, DegAE pretraining can still bring improvement. The visual effects are portrayed in Fig. 4, for the model trained from scratch, there are lots of rain residuals in the output images, while pretraining can help remove the rain more thoroughly.

5.3. Experiment on Image Motion Deblur

The DegAE pretraining can also bring considerable improvement on motion deblur task. We adopt GoPro [47] dataset for training and testing. It consists of 2,103 image pairs for training and 1,111 pairs for testing. Besides, we also test the model on HIDE [56] dataset. We report the results of existing methods for reference: DeblurGAN [29], DeblurGAN-v2 [30], SRN [60], SPAIR [49], HINet [6],

Table 3. Image derain results on benchmark datasets. DegAE pretraining can bring improvements up to 1.27dB for SwinIR backbone.

Method	Rain100L		Ran100H		Test100		Test1200		Test2800	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM	PSNR(dB)	SSIM	PSNR(dB)	SSIM	PSNR(dB)	SSIM
DerainNet [16]	27.03	0.884	14.92	0.592	22.77	0.810	23.38	0.835	24.31	0.861
RESCAN [35]	29.80	0.881	26.36	0.786	25.00	0.835	30.51	0.882	31.29	0.904
PreNet [53]	32.44	0.950	26.77	0.858	24.81	0.851	31.36	0.911	31.75	0.916
MSPFN [25]	32.40	0.933	28.66	0.860	27.50	0.876	32.39	0.916	32.82	0.930
MPRNet [72]	36.40	0.965	30.41	0.890	30.27	0.897	32.91	0.916	33.64	0.938
SwinIR	35.68	0.962	30.02	0.888	29.43	0.897	30.36	0.904	33.39	0.937
DegAE (SwinIR)	**36.95**	**0.969**	**30.10**	**0.891**	**30.16**	**0.902**	**30.53**	**0.905**	**33.48**	**0.938**
Uformer	36.26	0.968	27.01	0.884	28.19	0.902	32.09	0.904	27.36	0.916
DegAE (Uformer)	**36.80**	**0.970**	**27.47**	**0.885**	**28.73**	**0.902**	**32.17**	**0.908**	**27.44**	**0.917**
Restormer	38.38	0.975	32.19	0.911	31.65	0.924	32.88	0.923	33.61	0.943
DegAE (Restormer)	**38.83**	**0.977**	**32.19**	**0.911**	**31.77**	**0.924**	**32.99**	**0.925**	**33.66**	**0.944**

Table 4. Image motion deblur results (PSNR) on GoPro dataset and HIDE dataset.

Method	GoPro	HIDE
DeblurGAN [29]	28.70	24.51
DeblurGAN-v2 [30]	29.55	26.61
SRN [60]	30.26	28.36
SPAIR [49]	32.06	30.29
HINet [6]	32.71	–
MPRNet [72]	32.66	30.96
IPT [4]	32.52	–
NAFNet [5]	32.85	–
SwinIR	31.43	29.15
DegAE (SwinIR)	**31.90 (+0.47)**	**29.60 (+0.45)**
Restormer	32.60	31.10
DegAE (Restormer)	**33.03 (+0.43)**	**31.43 (+0.33)**
Uformer	33.04	30.92
DegAE (Uformer)	**33.16 (+0.12)**	**31.00 (+0.08)**

MPRNet [72], IPT [4], NAFNet [5]. Note that, for all compared methods in this paper, we do not apply Test-time Local Converter (TLC) proposed in [8] to improve test-time performance. The third row of Fig. 3 shows an example.

The quantitative results of motion deblur are shown in Tab. 4. By introducing DegAE pretraining, SwinIR achieves 0.47dB and 0.45dB improvement on GoPro and HIDE test set. Restormer yields a 0.43dB improvement on GoPro test set. In addition, compared with other methods, Uformer trained from scratch has already achieved the best performance, while DegAE pretraining can further enhance its performance, making it a new state-of-the-art model.

6. Conclusion

In this paper, we provide a comprehensive review of current pretraining methods for both high-level and low-level vision tasks. We categorize low-level vision tasks into low-cost task and high-cost task based on the difficulty of data acquisition. We claim that the pretrain-finetune scheme should prioritize high-cost downstream tasks. We introduce a new pretraining paradigm for low-level vision, called degradation autoencoder (DegAE). This approach effectively extracts informative representations that lead to significant improvements in model performance across various downstream tasks.

Figure 5. Visual comparison of training from scratch and the proposed DegAE pretraining on motion deblur effects.

Limitation. Although we have successfully validated the effectiveness of our design on several downstream tasks, there are myriad low-level vision tasks to explore. As the first general low-level vision pretraining paradigm, it can be further optimized. More effective pretraining solutions tailored to low-level vision are expected to emerge.

Acknowledgements. This work is partially supported by the National Key R&D Program of China (NO. 2022ZD0160100), and in part by Shanghai Committee of Science and Technology (Grant No. 21DZ1100100). This work is also supported in part by the National Natural Science Foundation of China under Grant (62276251), the Joint Lab of CAS -HK, and in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences (No. 2020356).

References

[1] 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 workshops*, pages 126–135, 2017. 3, 7

[2] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[4] 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. 1, 2, 3, 4, 5, 8

[5] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022. 2, 8

[6] 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. 7, 8

[7] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. *arXiv preprint arXiv:2205.04437*, 2022. 1, 2, 3, 4

[8] X Chu, L Chen, C Chen, and X Lu. Improving image restoration by revisiting global information aggregation. *arXiv preprint arXiv:2112.04491*, 5, 2021. 8

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[10] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 3

[11] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1, 2

[12] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2157–2167, 2020. 6, 7

[13] Jiangxin Dong and Jinshan Pan. Physics-based feature dehazing networks. In *European Conference on Computer Vision*, pages 188–204. Springer, 2020. 6, 7

[14] Yu Dong, Yihao Liu, He Zhang, Shifeng Chen, and Yu Qiao. Fd-gan: Generative adversarial networks with fusion-discriminator for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10729–10736, 2020. 2, 4

[15] 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. 1, 2, 3

[16] Xueyang Fu, Jiabin Huang, Xinghao Ding, Yinghao Liao, and John Paisley. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Transactions on Image Processing*, 26(6):2944–2956, 2017. 2, 7, 8

[17] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3855–3863, 2017. 7

[18] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2, 3

[19] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. Conditional sequential modulation for efficient global image retouching. In *European Conference on Computer Vision*, pages 679–695. Springer, 2020. 5

[20] 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. 1, 2, 3, 4

[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. 1, 2, 3, 4

[22] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010. 6, 7

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 5

[24] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8022–8031, 2019. 4

[25] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8346–8355, 2020. 2, 7, 8

[26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[27] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 1, 2

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[29] 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*, pages 8183–8192, 2018. 2, 4, 7, 8

[30] 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*, pages 8878–8887, 2019. 2, 7, 8

[31] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1, 5

[32] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018. 4, 7

[33] Dasong Li, Yi Zhang, Ka Chun Cheung, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Learning degradation representations for image deblurring. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 4

[34] Wenbo Li, Xin Lu, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 1, 2, 3, 4

[35] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 254–269, 2018. 7, 8

[36] Yu Li, Robby T Tan, Xiaojie Guo, Jiangbo Lu, and Michael S Brown. Rain streak removal using layer priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2736–2744, 2016. 4, 7

[37] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 4

[38] 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. 2, 3, 4, 5

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[40] Jiaying Liu, Wenhan Yang, Shuai Yang, and Zongming Guo. Erase or fill? deep joint recurrent rain removal and reconstruction in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3233–3242, 2018. 4

[41] Lin Liu, Lingxi Xie, Xiaopeng Zhang, Shanxin Yuan, Xiangyu Chen, Wengang Zhou, Houqiang Li, and Qi Tian. Tape: Task-agnostic prior embedding for image restoration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 447–464. Springer, 2022. 2

[42] 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*, pages 7314–7323, 2019. 6, 7

[43] Yihao Liu, Anran Liu, Jinjin Gu, Zhipeng Zhang, Wenhao Wu, Yu Qiao, and Chao Dong. Discovering" semantics" in super-resolution networks. *arXiv preprint arXiv:2108.00406*, 2021. 5

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[45] Michael R Lomnitz. Diffjpeg. `https://github.com/mlomnitz/DiffJPEG`, 2021. 5

[46] Yu Luo, Yong Xu, and Hui Ji. Removing rain from a single image via discriminative sparse coding. In *Proceedings of the IEEE international conference on computer vision*, pages 3397–3405, 2015. 4, 7

[47] 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*, pages 3883–3891, 2017. 4, 7

[48] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2, 3

[49] Kuldeep Purohit, Maitreya Suin, AN Rajagopalan, and Vishnu Naresh Boddeti. Spatially-adaptive image restoration using distortion-guided networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2309–2319, 2021. 2, 7, 8

[50] 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*, volume 34, pages 11908–11915, 2020. 6, 7

[51] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 1, 2

[52] 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. 2

[53] 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*, pages 3937–3946, 2019. 2, 7, 8

[54] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *European conference on computer vision*, pages 154–169. Springer, 2016. 2

[55] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3253–3261, 2018. 6, 7

[56] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5572–5581, 2019. 7

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[58] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *arXiv preprint arXiv:2204.03883*, 2022. 6, 7

[59] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 2

[60] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8174–8182, 2018. 2, 7, 8

[61] 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 workshops*, pages 114–125, 2017. 3, 7

[62] 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. 2

[63] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1905–1914, October 2021. 4

[64] 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*, pages 17683–17693, 2022. 2, 3, 5, 7

[65] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10551–10560, 2021. 6, 7

[66] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 1, 2, 3

[67] 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. 1, 2, 3, 4

[68] Wenhan Yang, Jiaying Liu, and Jiashi Feng. Frame-consistent recurrent video deraining with dual-level flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1661–1670, 2019. 4

[69] 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. 7

[70] Wenhan Yang, Robby T Tan, Shiqi Wang, Yuming Fang, and Jiaying Liu. Single image deraining: From model-based to data-driven and beyond. *IEEE Transactions on pattern analysis and machine intelligence*, 43(11):4059–4077, 2020. 2, 4, 7

[71] 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*, pages 5728–5739, 2022. 2, 3, 5, 7

[72] 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*, pages 14821–14831, 2021. 7, 8

[73] He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 695–704, 2018. 7

[74] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *IEEE transactions on circuits and systems for video technology*, 30(11):3943–3956, 2019. 7

[75] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 4

[76] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2737–2746, 2020. 4

[77] 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. 1, 2

[78] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 1, 2

[79] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 3

[80] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 3

[81] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 6