DegAE: A New Pretraining Paradigm for Low-level Vision

Yihao Liu1,2,3  Jingwen He1  Jinjin Gu1,4  Xiangtao Kong1,2,3  Yu Qiao1,2  Chao Dong2,1*

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.

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 $D_1$ and a reference image with degradation $D_2$. It attempts to transfer the degradation $D_2$ of the reference image to the input image, obtaining an output image with input image content, but with degradation $D_2$, as described in Fig. 1. Through such a learning paradigm, the model is expected to learn both
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 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 naïvely 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 denoising, 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 ×4 SR brings slight improvement (around 0.1dB). EDT [34] proposes a multi-related-task pretraining method that handles several highly related tasks, such as ×2, ×3, ×4 SR, on a partial ImageNet (200k) dataset. Each sub-task (e.g., ×4 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

1Rain model is single and fixed.
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 $I_{D_1}$ with degradation $D_1$ and a reference image $I_{D_2}^{ref}$ with degradation $D_2$. It aims to transfer the degradation $D_2$ to the input image, for obtaining an output image $I_D^{D_2}$ with input image content, but with reference degradation $D_2$. 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 $k$. After that, noise $n$ is added. Then JPEG compression with quality $q$ is applied. Specifically, we have $I = (I \otimes k + n)_{JPEG_q}$, where $\bar{I}$ is the degraded image. Following [63, 75], in terms of the choices of blur kernel $k$, we mainly consider isotropic and anisotropic Gaussian filters. For noise $n$, 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 $\phi$. 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.

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 $\phi$ that contains a pretrained SR-GAN [31] model and several downsampling layers to produce the degradation embeddings based on the given degraded reference images. Formally, given the degraded input image $I_{D1}$ and the degraded reference image $I_{D2}$, we have $I^D_2 = D(E(I_{D1}), \phi(I_{D1}^{ref}))$, where $I^D_2$ is the output image, which is expected to be close to the target image $I_{D2}$. Note that the reference $I_{D2}^{ref}$ 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 $L_{\text{content}}$, perceptual loss $L_{\text{per}}$, adversarial loss $L_{\text{adv}}$, and embedding loss $L_{\text{embed}}$.

Content Reconstruction Loss: For content consistency, we apply a simple Gaussian blur kernel $k$ 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: $L_{\text{content}} = \|I^D_2 \otimes k, \hat{I}^D_2 \otimes k\|^2$.

Adversarial Loss: We use generative adversarial learning...
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).

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.

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

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

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 $\beta_1 = 0.9$ and $\beta_2 = 0.99$ 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 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Rain100L</th>
<th>Ran100H</th>
<th>Test100</th>
<th>Test1200</th>
<th>Test2800</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR(dB)</td>
<td>SSIM</td>
<td>PSNR(dB)</td>
<td>SSIM</td>
<td>PSNR(dB)</td>
</tr>
<tr>
<td>DerainNet [16]</td>
<td>27.03</td>
<td>0.884</td>
<td>14.92</td>
<td>0.592</td>
<td>22.77</td>
</tr>
<tr>
<td>RESCAN [35]</td>
<td>29.80</td>
<td>0.881</td>
<td>26.36</td>
<td>0.786</td>
<td>25.00</td>
</tr>
<tr>
<td>PreNet [53]</td>
<td>32.44</td>
<td>0.950</td>
<td>26.77</td>
<td>0.858</td>
<td>24.81</td>
</tr>
<tr>
<td>MSPFN [25]</td>
<td>32.40</td>
<td>0.933</td>
<td>28.66</td>
<td>0.860</td>
<td>27.50</td>
</tr>
<tr>
<td>MPRNet [72]</td>
<td>36.40</td>
<td>0.965</td>
<td>30.41</td>
<td>0.890</td>
<td>30.27</td>
</tr>
<tr>
<td>SwinIR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DegAE (SwinIR)</td>
<td>36.95</td>
<td>0.969</td>
<td>30.10</td>
<td>0.891</td>
<td>30.16</td>
</tr>
<tr>
<td>UFormer</td>
<td>36.26</td>
<td>0.968</td>
<td>27.01</td>
<td>0.884</td>
<td>28.19</td>
</tr>
<tr>
<td>DegAE (Uformer)</td>
<td>36.80</td>
<td>0.970</td>
<td>27.47</td>
<td>0.885</td>
<td>28.73</td>
</tr>
<tr>
<td>Restormer</td>
<td>38.38</td>
<td>0.975</td>
<td>32.19</td>
<td>0.911</td>
<td>31.65</td>
</tr>
<tr>
<td>DegAE (Restormer)</td>
<td>38.83</td>
<td>0.977</td>
<td>32.19</td>
<td>0.911</td>
<td>31.77</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>GoPro</th>
<th>HIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeblurGAN [29]</td>
<td>28.70</td>
<td>24.51</td>
</tr>
<tr>
<td>SRN [60]</td>
<td>30.26</td>
<td>28.36</td>
</tr>
<tr>
<td>SPAIR [49]</td>
<td>32.06</td>
<td>30.29</td>
</tr>
<tr>
<td>HINet [6]</td>
<td>32.71</td>
<td>–</td>
</tr>
<tr>
<td>MPRNet [72]</td>
<td>32.66</td>
<td>30.96</td>
</tr>
<tr>
<td>IPT [4]</td>
<td>32.52</td>
<td>–</td>
</tr>
<tr>
<td>NAFNet [5]</td>
<td>32.85</td>
<td>–</td>
</tr>
<tr>
<td>SwinIR</td>
<td>31.43</td>
<td>29.15</td>
</tr>
<tr>
<td>DegAE (SwinIR)</td>
<td>31.90 (+0.47)</td>
<td>29.60 (+0.45)</td>
</tr>
<tr>
<td>Restormer</td>
<td>32.60</td>
<td>31.10</td>
</tr>
<tr>
<td>DegAE (Restormer)</td>
<td>33.03 (+0.43)</td>
<td>31.43 (+0.33)</td>
</tr>
<tr>
<td>UFormer</td>
<td>33.04</td>
<td>30.92</td>
</tr>
<tr>
<td>DegAE (Uformer)</td>
<td>33.16 (+0.12)</td>
<td>31.00 (+0.08)</td>
</tr>
</tbody>
</table>

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.

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


[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


[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


23300

