# DegAE: A New Pretraining Paradigm for Low-level Vision
# – Supplementary File –

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

*In this supplementary file, we provide more supporting materials. First, we introduce the designing philosophy of DegAE. Second, we present the experimental results on low-cost tasks, including Gaussian denoise and super-resolution. Third, we conduct ablation studies on the pretraining losses. Then, we describe more implementation details, such as the architecture of the decoder, training strategy,* etc. *Last, more visual results are presented.*

## 1. The Philosophy of DegAE

The purpose of low-level vision is to produce natural clean images. To achieve this, the model is expected to learn a good and general representation of natural images. However, previous literatures have shown that deep networks tend to overfit the training degradation rather than actually learn the distribution of natural images [12]. In the design of DegAE, the encoder extract features from various degraded input images and the decoder tries to transfer another degradation to the input degraded images. Therefore, our method implicitly has two stages: restore the degraded image to a clean image, and then add new degradation to the clean image. This suggests that the encoder has to project all degraded images into a unified distribution of clean images. To verify this, we train the encoder-decoder structure of DegAE with different objectives, *i.e.*, SR, denoise, multi-task restoration (MTR) and DegAE. To be specific, SR refers to $\times 4$ classical super-resolution; denoise refers to Gaussian denoise with noise level [0, 50]; multi-degradation restoration includes various degradation settings mentioned in [22]; DegAE means our proposed pretext task. Then, following [13], we convert and visualize the encoder's output feature distributions of different input degradations. The PIES dataset is borrowed from [13], which includes patch-based images with various degradations. Each degradation contains 800 images. As shown in Fig. 1, the encoder of DegAE successfully transfer various degradations into similar distributions, while other training schemes will cause large difference in the encoder's distribution. This verifies our hypothesis that the encoder has successfully pull various degradations into a unified clean image space. Interestingly, we find that the distributions of MTR will become unanimous until the last output layers. The encoder's output distributions are still separated for different degradations. On the contrary, our DegAE can effectively project different degradations into a unified distribution at the encoder stage.

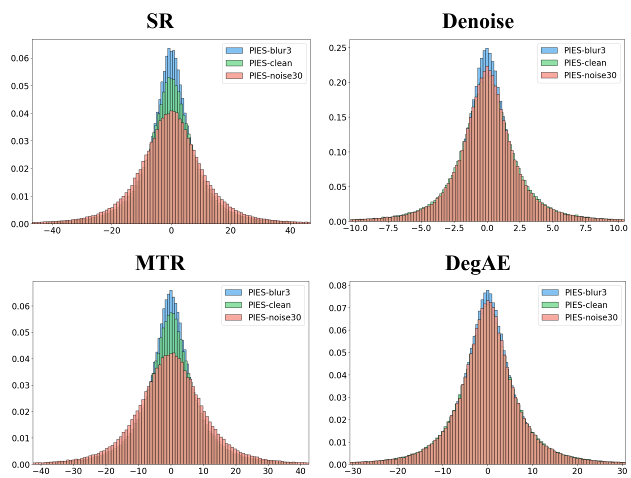



Figure 1. Feature distributions of encoder with different training schemes. The encoder of DegAE can successfully transfer variosu degradaed input into a unanimous distribution.

## 2. Experiment on Low-cost Tasks

In addition with high-cost tasks, we also perform experiments on several low-cost tasks, like image super-resolution and Gaussian denoise.

**Image Super-resolution.** We also conduct finetuning on super-resolution (SR), which is a classical low-level vision task. Classical SR task [4, 30] assumes that the im-

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

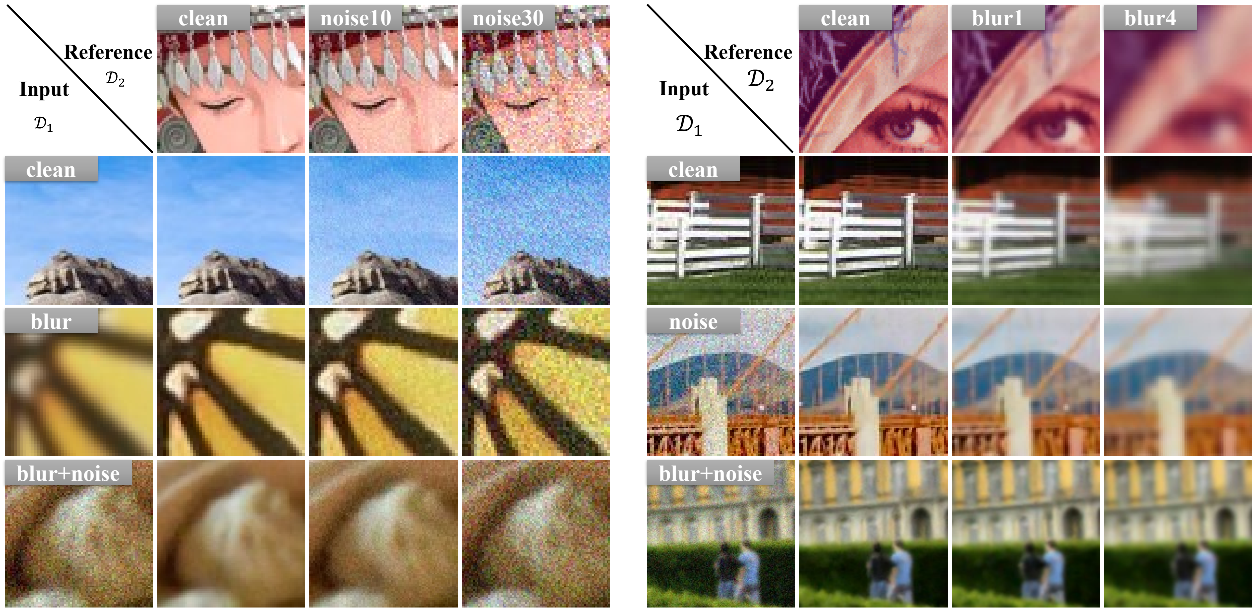



Figure 2. Example results of DegAE pretraining.

age downsampling process is modeled by bicubic downsampling. Different from previous downstream tasks, the training pairs of classical SR can be synthesized on-the-fly, thus the amount of data can be regarded as unlimited. We consider the $\times 4$ SR task and adopt DF2K dataset (DIV2K [1]+Flickr2K [21]) to synthesize the training pairs. The trained models are evalated on Set5 [2], Set14 [26], BSDS100 [16] and Urban100 [8] datasets. For reference, we also report the performance of several state-of-the-art methods: RCAN [30], SAN [3], HAN [19], NLSA [18]. We calculate the PSNR and SSIM scores on the Y channel in YCbCr color space.

From Tab. 1, it can be seen that pretraining does not bring much improvement on SR task. For example, SwinIR only gains 0.03dB improvement at most on Set5 and Urban100 dataset. The performance gain of Uformer and Restormer is also marginal (less than 0.1dB). This is reasonable since the image pairs of SR can be obtained infinitely during training and the degradation process (bicubic downsampling) is fixed as well. Unlimited training data weakens the significance of pretraining, as sufficient data can facilitate full training of the model.

**Image Gaussian Denoising.** We further perform denoising experiments with additive white Gaussian noise. Similar to SR task, the Gaussian noise is synthesized on-the-fly during training. For better universality, we train SwinIR model with noise level sampled from a wide range of $[0, 50]$, rather than training on a certain single noise level. Besides, we also retrain several baseline models for reference, including DnCNN [28], IRCNN [29], DRUNet [27] and modified SRResNet [23]. We then test the trained models on Kodak24 [5], CBSD68 [17] and Urban100 [8] datasets with different noise levels.

As shown in Tab. 2, with DegAE pretraining, SwinIR achieves the largest improvement of 0.82dB on CBSD68 dataset with noise level 15. However, on Kodak24 and Urban100 datasets, the improvement is relatively small. Especially for Kodak24 set, the improvement is very marginal. Note again, we can synthesize the Gaussian noise data during training process almost without cost. For such tasks with unlimited amounts of data, as long as the original background images are sufficient, the model can already learn good enough representations without the additional power of pretraining.

## 3. Influence of Pretraining Losses

In training DegAE, we employ four loss functions, namely content reconstruction loss $\mathcal{L}_{content}$, perceptual loss $\mathcal{L}_{per}$, adversarial loss $\mathcal{L}_{adv}$, and embedding loss $\mathcal{L}_{embed}$. These losses are commonly utilized in GAN-based superresolution (SR) methods. Our ablation studies reveal that the GAN loss and perceptual loss are crucial for effectively learning complex degradations, while the content reconstruction loss helps to preserve image contents. As illustrated in Fig. 3, when solely $\mathcal{L}_{content}$ loss ($L_1$ loss) is used, the model fails to generate any noise degradation in the output images. In the case of only adopting adversarial loss $\mathcal{L}_{adv}$, the model suffers from model collapse, which prevents it from generating normal images. Similarly, if solely $\mathcal{L}_{per}$ loss is used, the produced noise degradation lacks fi-

| | Set5 | | Set14 | | BSDS100 | | Urban100 | |
|---|---|---|---|---|---|---|---|---|
| **Method** | PSNR(dB) | SSIM | PSNR(dB) | SSIM | PSNR(dB) | SSIM | PSNR(dB) | SSIM |
| RCAN [30] | 32.63 | 0.900 | 28.87 | 0.789 | 27.77 | 0.744 | 26.82 | 0.809 |
| SAN [3] | 32.64 | 0.900 | 28.92 | 0.789 | 27.78 | 0.744 | 26.79 | 0.807 |
| HAN [19] | 32.64 | 0.900 | 28.90 | 0.789 | 27.80 | 0.744 | 26.85 | 0.809 |
| NLSA [18] | 32.59 | 0.900 | 28.87 | 0.789 | 27.78 | 0.744 | 26.96 | 0.811 |
| Uformer | 31.84 | 0.894 | 28.47 | 0.783 | 27.40 | 0.737 | 26.32 | 0.795 |
| **DegAE (Uformer)** | **31.88** | **0.894** | **28.50** | **0.784** | **27.42** | **0.737** | **26.32** | **0.795** |
| Restormer | 32.57 | 0.900 | 28.93 | 0.789 | 27.79 | 0.743 | 26.79 | 0.805 |
| **DegAE (Restormer)** | **32.62** | **0.900** | **28.99** | **0.790** | **27.80** | **0.744** | **26.82** | **0.806** |
| SwinIR | 32.73 | 0.902 | 29.07 | 0.793 | 27.88 | 0.747 | 27.32 | 0.821 |
| **DegAE (SwinIR)** | **32.76** | **0.902** | **29.08** | **0.793** | **27.89** | **0.747** | **27.35** | **0.822** |

Table 1. Image super-resolution results.

| | Kodak24 | | CBSD68 | | Urban100 | |
|---|---|---|---|---|---|---|
| **Method** | $\sigma$=15 | $\sigma$=25 | $\sigma$=15 | $\sigma$=25 | $\sigma$=15 | $\sigma$=25 |
| DnCNN [28] | 31.24 | 27.19 | 30.26 | 26.10 | 29.81 | 25.28 |
| IRCNN [29] | 31.37 | 27.33 | 30.37 | 26.25 | 29.93 | 25.44 |
| SRResNet [23] | 32.00 | 27.98 | 30.83 | 26.72 | 31.02 | 26.40 |
| DRUNet [27] | 32.18 | 28.13 | 30.72 | 26.48 | 31.17 | 26.53 |
| SwinIR | 32.13 | 28.20 | 30.03 | 26.48 | 31.30 | 26.74 |
| **DegAE (SwinIR)** | **32.18** | **28.30** | **30.56** | **26.80** | **31.42** | **26.85** |

Table 2. Image Gaussian denoising results (PSNR) on test datasets.

delity and does not effectively transfer different levels of blur degradation to the noise input. In summary, incorporating multiple loss functions is essential for successful training of the DegAE model.

# 4. Implementation Details

## 4.1. Details on Degradation Autoencoder

**Detailed Structure of Decoder** For pretraining, the decoder is a pure CNN architecture that contains four residual blocks [7]. For each residual block, a degradation injection module is introduced to modulate the intermediate features. Specifically, the degradation injection module accepts a degradation embedding and then outputs the modulators–scaling $\alpha$ and shifting $\beta$ parameters using two independent fully connected layers for global feature modulation (GFM) [6]. The formulation of GFM is given by:

$$GFM(x_i) = \alpha * x_i + \beta, \tag{1}$$

where $x_i \in \mathbb{R}^{C\times H\times W}$ is the intermediate feature map. $C$, $H$ and $W$ are channels, height and width, respectively. In order to generate noisy images, we also introduce random noise map and the corresponding weighting parameter $w$ that is learned during training. This noise injection is performed after global feature modulation: $\tilde{x}_i = GFM(x_i) + w * \eta$, where $\eta \sim N(0,1)$.

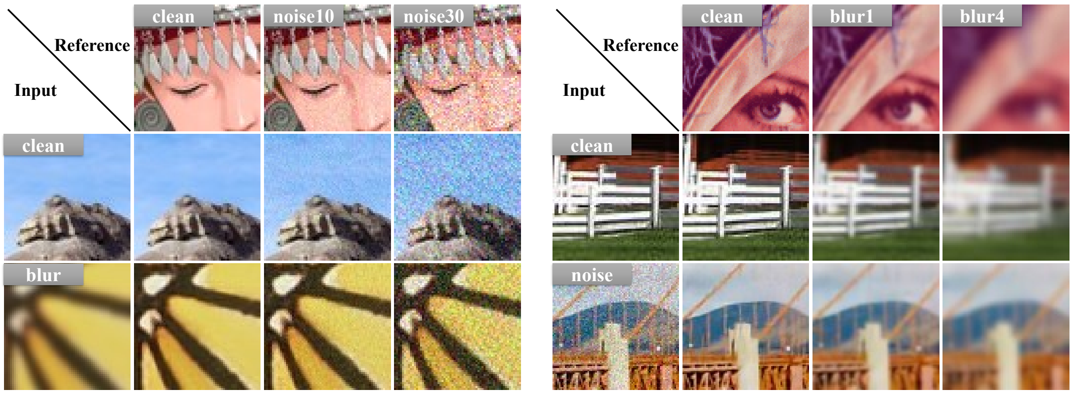
The degradation embedding is produced by a degradation representor $\phi$ based on the given reference degraded image $I_{ref}^{\mathcal{D}_2}$ with degradation $\mathcal{D}_2$. The degradation representor $\phi$ contains a pretrained SRGAN [10] model, three convolution layers with stride 2 for downsampling, one global average pooling layer, and three FC+LeakyReLU [15] layers. The channel dimension of the degradation embedding is 512. The extracted degradation embedding will be fed into the GFM to modulate the intermediate features of the decoder, which governs the generation of different degradations. In practice, we find this simple design works well, especially for simulating blur and noise degradations. Nevertheless, better design could be explored for further work.

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. For super resolution, we additionally add some convolution layers and pixelshuffle [20] layers.

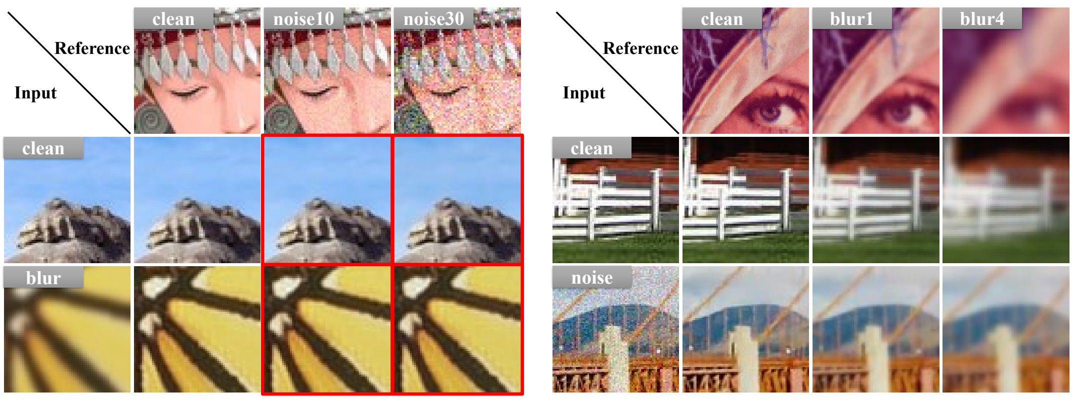
**Details of Degradation Input.** As for blur operation, we use Gaussian kernels, generalized Gaussian kernels and plateau-shaped kernels and their probabilities are 0.7, 0.15, 0.15, respectively. The kernel size is selected from 7, 9, ...21 randomly. For generalized Gaussian and plateau-shaped kernels, the shape parameter $\beta$ is sampled from [0.5, 4] and [1, 2], respectively. The probability of sinc kernel is set to 0.1. The Gaussian noises and Poisson noises are employed with probability 0.5. We set the noise sigma range and Poisson noise scale to [1, 30] and [0.05, 3], respectively The gray noise probability is set to 0.4. JPEG compression quality factor is set to [30, 95]. The final sinc filter is applied with a probability of 0.8.

## 4.2. Details on Finetuning Strategy

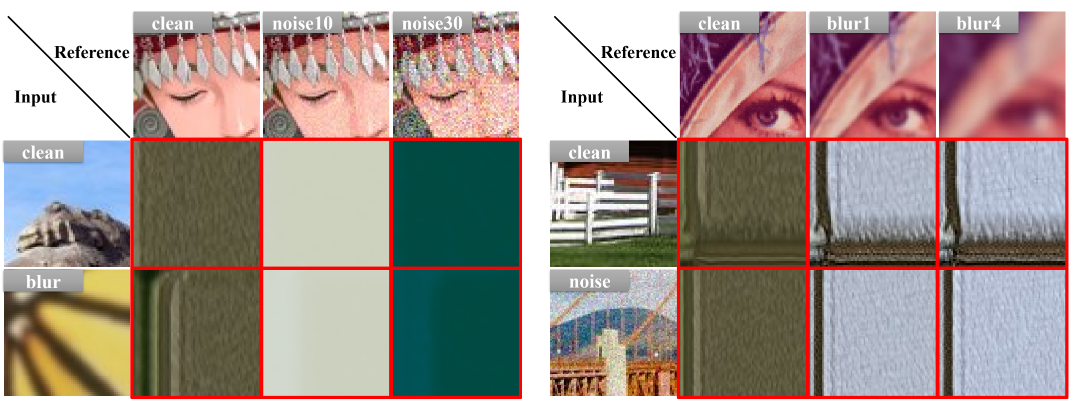
For finetuning, we replace the decoder with a single convolution layer. The kernel size is $3 \times 3$ and the output channel is 3. The parameters of the backbone are initialized from DegAE pretraining. The initial learning rate is 3e-4 and is cosine decayed to 1e-6. We randomly augment



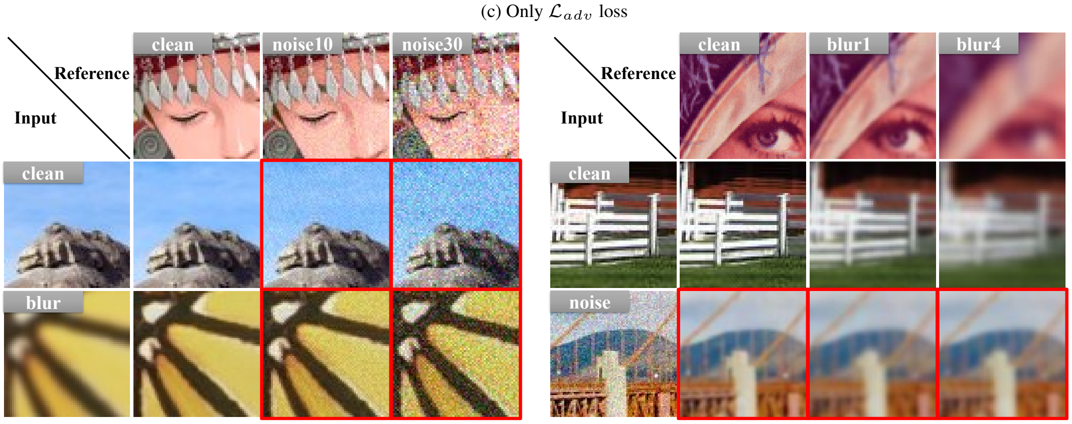
(a) Full losses



(b) Only $\mathcal{L}_{content}$ loss



(c) Only $\mathcal{L}_{adv}$ loss



(b) Only $\mathcal{L}_{per}$ loss

Figure 3. Effect of different losses. The imperfect cases are highlighted in red rectangles.

the training samples using the horizontal flipping and rotate the images by $90^\circ$, $180^\circ$, and $270^\circ$. For all downstream tasks, we adopt $L_1$ loss. For SwinIR [11] backbone, the input patch size is $128 \times 128$ [1]. The Adam optimizer [9] is adopted as the original SwinIR paper. For Uformer [24] and Restormer [25] backbone, the input patch size is $256 \times 256$. The AdamW [14] optimizer is adopted.

# 5. Additional Visual Results

## 5.1. Visual Results of Downstream Tasks

We provide more visual results of downstream tasks in Fig. 4, including image dehaze, derain and motion deblur.

## 5.2. Effects of DegAE Pretraining-Finetuning

As shown in Fig. 6, 7, and 8, DegAE pretraining can reduce the generated artifacts and help remove the haze/rain/blur more thoroughly, compared to training from scratch.

## 5.3. Visual Comparison with MAE

In Fig. 5, we show some visual examples of ViT, MAE, FFA-Net, Uformer and DegAE on dehaze dataset. ViT-based pure Transformer architecture is not friendly to low-level vision tasks, due to the rough patch-splitting strategy. The produced visual results contain much box artifacts. In addition, MAE pretraining does not bring effective improvement. FFA-Net is a pure CNN-based model and Uformer contains CNN pre-processing and post-processing as well. Their results do not contain artifacts as ViT. This implies that CNN structure has its unique advantages for low-level vision tasks. Further, by adopting the proposed DegAE pretraining scheme, Uformer achieve significant improvement. This clearly shows the effectiveness of DegAE, which is tailored to low-level vision.

[1] Since the architecture of SwinIR costs lots of GPU memory, we do not set the patch size to $256 \times 256$ as other backbones.

# 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. 2

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 2

[3] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019. 2, 3

[4] 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

[5] Rich Franzen. Kodak lossless true color image suite. http://r0k.us/graphics/kodak/, 1999. 2

[6] 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. 3

[7] 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. 3

[8] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015. 2

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[10] 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. 3

[11] 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. 5

[12] 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. 1

[13] Yihao Liu, Hengyuan Zhao, Jinjin Gu, Yu Qiao, and Chao Dong. Evaluating the generalization ability of super-resolution networks. *arXiv preprint arXiv:2205.07019*, 2022. 1

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[15] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, Georgia, USA, 2013. 3

[16] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001. 2

[17] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE*


Figure 4. 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.

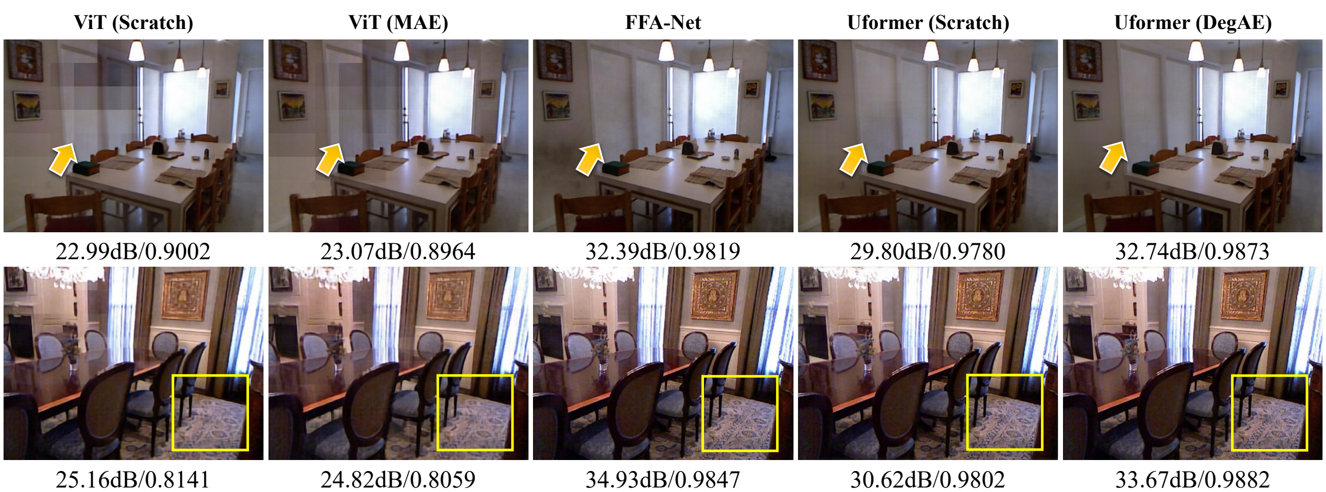



Figure 5. Visual comparison with ViT, MAE, FFA-Net, Uformer and DegAE.

*International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001. 2

[18] 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*, pages 3517–3526, 2021. 2, 3

[19] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European conference on computer vision*, pages 191–207. Springer, 2020. 2, 3

[20] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 3

[21] 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 *Proceed-*

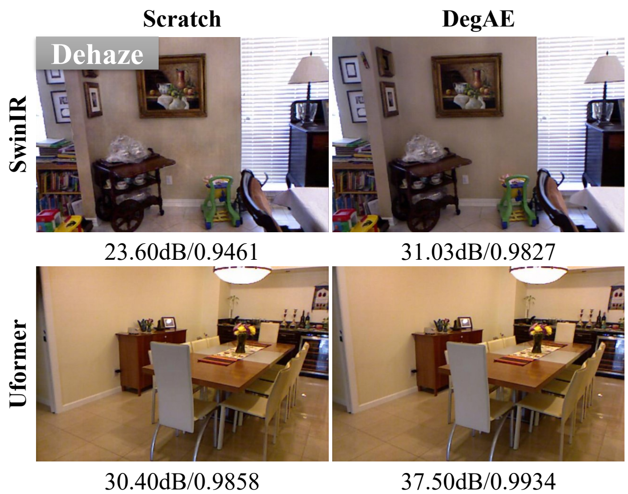



Figure 6. Visual comparison of training from scratch and DegAE pretraining on dehaze effects.

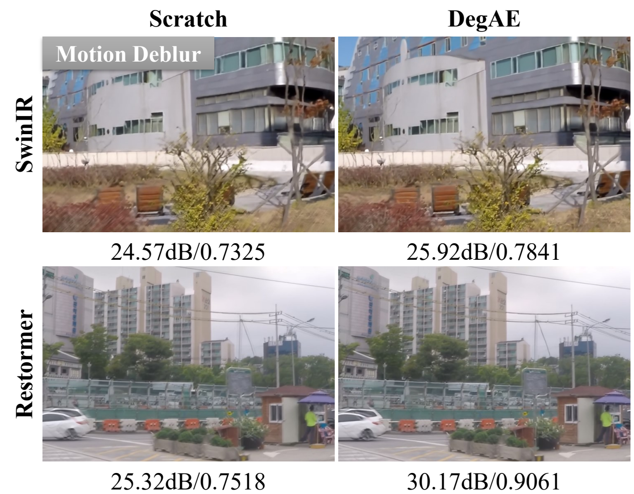



Figure 8. Visual comparison of training from scratch and DegAE pretraining on motion deblur effects.

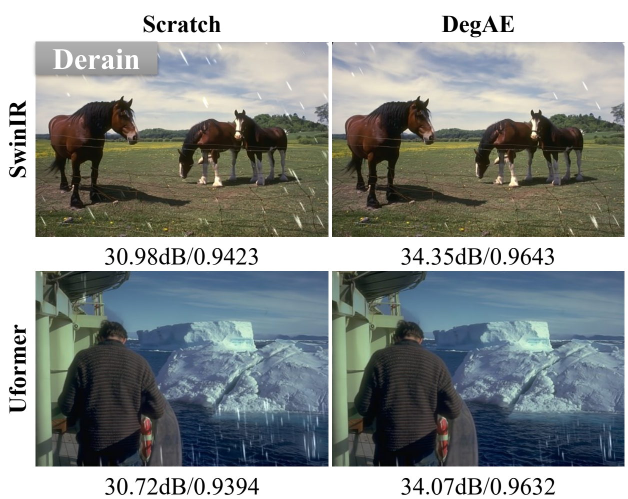



Figure 7. Visual comparison of training from scratch and DegAE pretraining on derain effects.

*ings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017. 2

[22] 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. 1

[23] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 63–79, Cham, 2019. Springer International Publishing. 2, 3

[24] 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. 5

[25] 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. 5

[26] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010. 2

[27] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 3

[28] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 2, 3

[29] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017. 2, 3

[30] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 1, 2, 3