Pre-Trained Image Processing Transformer

Hanting Chen[1,2], Yunhe Wang[2]*, Tianyu Guo[1,2], Chang Xu[3], Yiping Deng[4],
Zhenhua Liu[2,5,6], Siwei Ma[5,6], Chunjing Xu[2], Chao Xu[1], Wen Gao[5,6]

[1] Key Lab of Machine Perception (MOE), Dept. of Machine Intelligence, Peking University. [2] Noah's Ark Lab, Huawei Technologies.
[3] School of Computer Science, Faculty of Engineering, The University of Sydney. [4] Central Software Institution, Huawei Technologies.
[5] Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University. [6] Peng Cheng Laboratory.

htchen@pku.edu.cn, yunhe.wang@huawei.com

Abstract

*As the computing power of modern hardware is increasing strongly, pre-trained deep learning models (*e.g.*, BERT, GPT-3) learned on large-scale datasets have shown their effectiveness over conventional methods. The big progress is mainly contributed to the representation ability of transformer and its variant architectures. In this paper, we study the low-level computer vision task (*e.g.*, denoising, super-resolution and deraining) and develop a new pre-trained model, namely, image processing transformer (IPT). To maximally excavate the capability of transformer, we present to utilize the well-known ImageNet benchmark for generating a large amount of corrupted image pairs. The IPT model is trained on these images with multi-heads and multi-tails. In addition, the contrastive learning is introduced for well adapting to different image processing tasks. The pre-trained model can therefore efficiently employed on desired task after fine-tuning. With only one pre-trained model, IPT outperforms the current state-of-the-art methods on various low-level benchmarks. Code is available at https://github.com/huawei-noah/Pretrained-IPT and https://gitee.com/mindspore/mindspore/tree/master/model_zoo/research/cv/IPT*

1. Introduction

Image processing is one component of the low-level part of a more global image analysis or computer vision system. Results from the image processing can largely influence the subsequent high-level part to perform recognition and understanding of the image data. Recently, deep learning has been widely applied to solve low-level vision tasks, such as image super-resolution, inpainting, deraining and colorization. As many image processing tasks are related, it is natural to expect a model pre-trained on one dataset can be helpful for another. But few studies have generalized pre-training across image processing tasks.

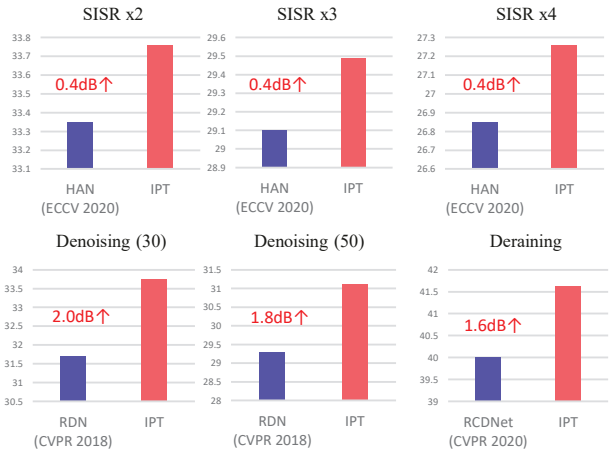

Figure 1. Comparison on the performance of the proposed IPT and the state-of-the-art image processing models on different tasks.

Pre-training has the potential to provide an attractive solution to image processing tasks by addressing the following two challenges: First, task-specific data can be limited. This problem is exacerbated in image processing task that involves the paid-for data or data privacy, such as medical images [8] and satellite images [73]. Various inconsistent factors (e.g. camera parameter, illumination and weather) can further perturb the distribution of the captured data for training. Second, it is unknown which type of image processing job will be requested until the test image is presented. We therefore have to prepare a series of image processing modules at hand. They have distinct aims, but some underlying operations could be shared.

It is now common to have pre-training in natural language processing and computer vision [12]. For example, the backbones of object detection models [86, 85] are often pre-trained on ImageNet classification [18]. A num-

*Corresponding author

ber of well-trained networks can now be easily obtained from the Internet, including AlexNet [41], VGGNet [56] and ResNet [33]. The seminal work Transformers [61] have been widely used in many natural language processing (NLP) tasks, such as translation [64] and question-answering [58]. The secret of its success is to pre-train transformer-based models on a large text corpus and fine-tune them on the task-specific dataset. Variants of Transformers, like BERT [19] and GPT-3 [5], further enriched the training data and improved the pre-training skills. There have been interesting attempts on extending the success of Transformers to the computer vision field. For example, Wang *et al*. [62] and Fu *et al*. [25] applied the self-attention based models to capture global information on images. Carion *et al*. [7] proposed DERT to use transformer architectures for an end-to-end object detection. Most recently, Dosovitskiy *et al*. [22] introduced Vision Transformer (ViT) to treat input images as 16×16 words and attained excellent results on image recognition.

The aforementioned pre-training in computer vision and natural language mostly investigate a pretest classification task, but both the input and the output in an image processing task are images. A straightforward application of these existing pre-training strategies might not be feasible. Further, how to effectively address different target image processing tasks in the pre-training stage remains a hard challenge. It is also instructive to note that the pre-training of image processing models enjoys a convenience of self-generating training instances based on the original real images. The synthetically manipulated images are taken for training, while the original image itself is the ground-truth to be reconstructed.

In this paper, we develop a pre-trained model for image processing using the transformer architecture, namely, Image Processing Transformer (IPT). As the pre-trained model needs to be compatible with different image processing tasks, including super-resolution, denoising, and deraining, the entire network is composed of multiple pairs of head and tail corresponding to different tasks and a single shared body. Since the potential of transformer needs to be excavated using large-scale dataset, we should prepair a great number of images with considerable diversity for training the IPT model. To this end, we select the ImageNet benchmark which contains various high-resolution with 1,000 categories. For each image in the ImageNet, we generate multiple corrupted counterparts using several carefully designed operations to serve different tasks. For example, training samples for the super-resolution task are generated by downsampling original images. The entired dataset we used for training IPT contains about over 10 millions of images.

Then, the transformer architecture is trained on the huge dataset as follows. The training images are input to the specific head, and the generated features are cropped into patches (*i.e*., "words") and flattened to sequences subsequently. The transformer body is employed to process the flattened features in which position and task embedding are utilized for encoder and decoder, respectively. In addition, tails are forced to predict the original images with different output sizes according to the specific task. Moreover, a contrastive loss on the relationship between patches of different inputs is introduced for well adopting to different image processing tasks. The proposed image processing transformer is learned in an end-to-end manner. Experimental results conducted on several benchmarks show that the pre-trained IPT model can surpass most of existing methods on their own tasks by a significant enhancement after fine-tuning.

2. Related Works

2.1. Image Processing

Image processing consists of the manipulation of images, including super-resolution, denoising, dehazing, deraining, debluring, etc. There are a variety of deep-learning-based methods proposed to conduct on one or many kinds of image processing tasks. For the super-resolution, Dong *et al*. propose SRCNN [20, 21] which are considered as pioneering works introducing end-to-end models that reconstructs HR images from their LR counterparts. Kim *et al*. [39] further explore the capacity of deep neural network with a more deeper convolutional network. Ahn *et al*. [2] and Lim *et al*. [47] propose introduce residual block into SR task. Zhang *et al*. [80] and Anwar and Barnes [3] utilize the power of attention to enhance the performance on SR task. A various excellent works are also proposed for the other tasks, such as denoising [60, 31, 36, 42, 24], dehazing [6, 43, 74, 71], deraining [35, 69, 55, 29, 65, 44], and debluring [59, 50, 23, 10]. Different from above methods, we dig the capacity of both big models and huge volume of data. Then a pre-training model handling several image processing tasks is introduced.

2.2. Transformer

Transformer [61] and its variants have proven its success being powerful unsupervised or self-supervised pre-training frameworks in various natural language processing tasks. For example, GPTs [52, 53, 5] are pre-trained in a autoregressive way that predicting next word in huge text datasets. BERT [19] learns from data without explicit supervision and predicts a masking word based on context. Colin *et al*. [54] proposes a universal pre-training framework for several downstream tasks. Yinhan *et al*. [49] proposes a robust variant for original BERT.

Due to the success of Transformer-based models in the NLP field, there are many attempts to explore the benefits of

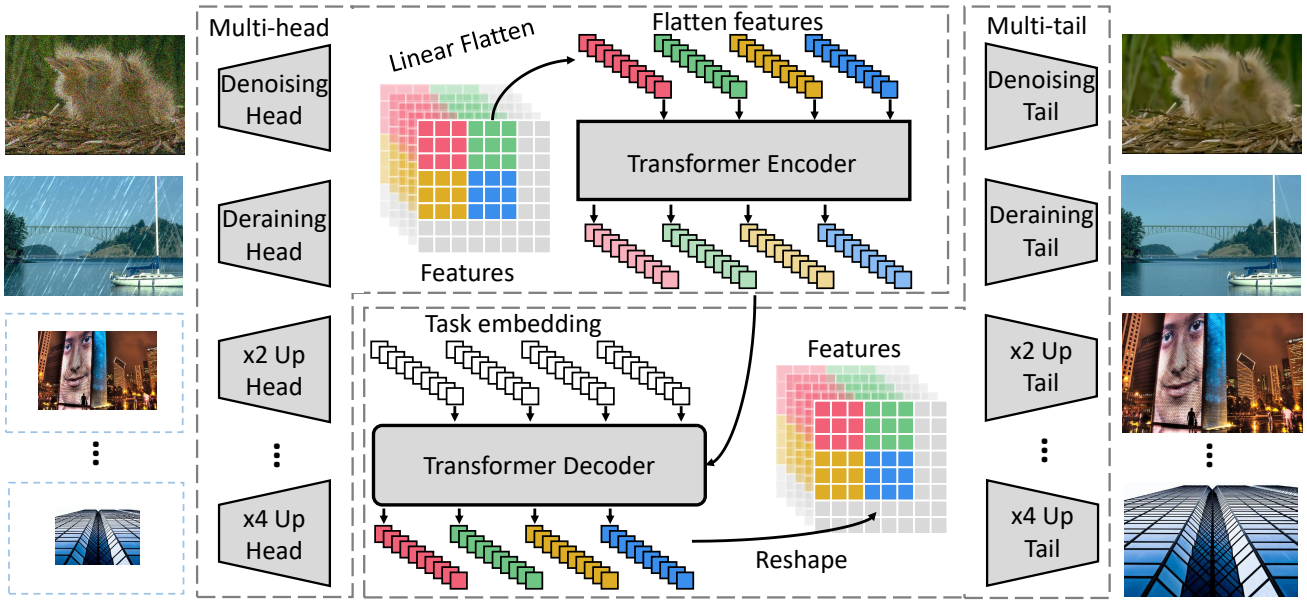

Figure 2. The diagram of the proposed image processing transformer (IPT). The IPT model consists of multi-head and multi-tail for different tasks and a shared transformer body including encoder and decoder. The input images are first converted to visual features and then divided into patches as visual words for subsequent processing. The resulting images with high visual quality are reconstructed by ensembling output patches.

Transformer in computer vision tasks. These attempts can be roughly divided into two types. The first is to introduce self-attention into the traditional convolutional neural network. Yuan *et al*. [72] introduce spatial attention for image segmentation. Fu *et al*. [26] proposes DANET utilizing the context information by combining spatial and channel attention. Wang *et al*. [66], Chen *et al*. [15], Jiang *et al*. [37] and Zhang *et al*. [79] also augment features by self-attention to enhance model performance on several high-level vision tasks. The other type is to replace convolutional neural network with self-attention block. For instance, Kolesnikov *et al*. [40] and Dosovitskiy [22] conduct image classification with transformer block. Carion *et al*. [7] and Zhu *et al*. [88] implement transformer-based models in detection. Chen *et al*. [11] proposes a pre-trained GPT model for generative and classification tasks. Wu *et al*. [68] and Zhao *et al*. [84] propose pre-training methods for teansformer-based models for image recognition task. Jiang *et al*. [38] propose the TransGAN to generate images using Transformer. However, few related works focus on low-level vision tasks. In this paper, we explore a universal pre-training approach for image processing tasks.

3. Image Processing Transformer

To excavate the potential use of transformer on image processing tasks for achieving better results, here we present the image processing transformer by pre-training on large-scale dataset.

3.1. IPT architecture

The overall architecture of our IPT consists of four components: heads for extracting features from the input corrupted images (*e.g*., images with noise and low-resolution images), an encoder-decoder transformer is established for recovering the missing information in input data, and tails are used formapping the features into restored images. Here we briefly introduce our architecture, details can be found in the supplementary material.

Heads. To adjust different image processing task, we use a multi-head architecture to deal with each task separately, where each head consists of three convolutional layers. Denote the input image as $x \in \mathbb{R}^{3\times H\times W}$ (3 means R, G, and B), the head generates a feature map $f_H \in \mathbb{R}^{C\times H\times W}$ with C channels and same height and width (typical we use $C = 64$). The calculation can be formulated as $f_H = H^i(x)$, where H^i $(i = \{1, \ldots, N_t\})$ denote the head for the ith task and N_t denotes the number of tasks.

Transformer encoder. Before input features into the transformer body, we split the given features into patches and each patch is regarded as a "word". Specifically, the features $f_H \in \mathbb{R}^{C\times H\times W}$ are reshaped into a sequence of patches, *i.e*., $f_{p_i} \in \mathbb{R}^{P^2\times C}, i = \{1, \ldots, N\}$, where $N = \frac{HW}{P^2}$ is the number of patches (*i.e*., the length of sequence) and P is patch size. To maintain the position information of each patch, we add learnable position encodings $E_{p_i} \in \mathbb{R}^{P^2\times C}$ for each patch of feature f_{p_i} following [22, 7], and $E_{p_i} + f_{p_i}$ will be directly input into the transformer encoder. The architecture of encoder layer is

following the original structure in [61], which has a multi-head self-attention module and a feed forward network. The output of encoder $f_{E_i} \in \mathbb{R}^{P^2 \times C}$ for each patch has the same size to that of the input patch f_{p_i}. The calculation can be formulated as

$$\begin{aligned} &y_0 = [E_{p_1} + f_{p_1}, E_{p_2} + f_{p_2}, \ldots, E_{p_N} + f_{p_N}], \\ &q_i = k_i = v_i = \text{LN}(y_{i-1}), \\ &y'_i = \text{MSA}(q_i, k_i, v_i) + y_{i-1}, \\ &y_i = \text{FFN}(\text{LN}(y'_i)) + y'_i, \qquad i = 1, \ldots, l \\ &[f_{E_1}, f_{E_2}, \ldots, f_{E_N}] = y_l, \end{aligned} \tag{1}$$

where l denotes the number of layers in the encoder, MSA denotes the multi-head self-attention module in the conventional transformer model [61], LN denotes the layer normalization [4] and FFN denotes the feed forward network, which contains two fully connected layers.

Transformer decoder. The decoder also follows the same architecture and takes the output of decoder as input in the transformer body, which consists of two multi-head self-attention (MSA) layers and one feed forward network (FFN). The difference to that of the original transformer here is that we utilize a task-specific embedding as an additional input of the decoder. These task-specific embeddings $E_t^i \in \mathbb{R}^{P^2 \times C}, i = \{1, \ldots, N_t\}$ are learned to decode features for different tasks. The calculation of decoder can be formulated as:

$$\begin{aligned} &z_0 = [f_{E_1}, f_{E_2}, \ldots, f_{E_N}], \\ &q_i = k_i = \text{LN}(z_{i-1}) + E_t, v_i = \text{LN}(z_{i-1}), \\ &z'_i = \text{MSA}(q_i, k_i, v_i) + z_{i-1}, \\ &q'_i = \text{LN}(z'_i) + E_t, k'_i = v'_i = \text{LN}(z_0), \\ &z''_i = \text{MSA}(q'_i, k'_i, v'_i) + z'_i, \\ &z_i = \text{FFN}(\text{LN}(z''_i)) + z''_i, \qquad i = 1, \ldots, l \\ &[f_{D_1}, f_{D_2}, \ldots, f_{D_N}] = y_l, \end{aligned} \tag{2}$$

where $f_{D_i} \in \mathbb{R}^{P^2 \times C}$ denotes the outputs of decoder. The decoded N patched features with size $P^2 \times C$ are then reshaped into the features f_D with size $C \times H \times W$.

Tails. The properties of tails are same as those of heads, we use multi tails to deal with different tasks. The calculation can be formulated as $f_T = T^i(f_D)$, where T^i $(i = \{1, \ldots, N_t\})$ denote the head for the ith task and N_t denotes the number of tasks. The output f_T is the resulted images size of $3 \times H' \times W'$ which is determined by the specific task. For example, $H' = 2H, W = 2W$ for a $2\times$ super-resolution task.

3.2. Pre-training on ImageNet

Besides the architecture of transformer itself, one of the key factors for successfully training an excellent transformer is that the well use of large-scale datasets. Compared with image classification, the number of available data used for image processing task is relatively small (*e.g.*, only 2000 images on DIV2K dataset for the image super-resolution task), we propose to utilize the well-known ImageNet as the baseline dataset for pre-training our IPT model, then we generate the entire dataset for several tasks (*e.g.*, super-resolution and denosing) as follows.

As the images in the ImageNet benchmark are of high diversity, which contains over 1 million of natural images from 1,000 different categories. These images have abundant texture and color information. We first remove the semantic label and manually synthesize a variety of corrupted images from these unlabeled images with a variety of degradation models for different tasks. Note that synthesized dataset is also usually used in these image processing tasks and we use the same degeneration methods as suggested in [30, 1]. For example, super-resolution tasks often take bicubic degradation to generate low-resolution images, denoising tasks add Gaussian noise in clean images with different noise level to generate the noisy images. These synthesized images can significantly improve the performance of learned deep networks including both CNN and transformer architectures, which will be shown in the experiment part. Basically, the corrupted images are synthesized as:

$$I_{corrupted} = \boldsymbol{f}(I_{clean}), \tag{3}$$

where $\boldsymbol{f}$ denotes the degradation transformation, which is depended on the specific task: for the super-resolution task, $\boldsymbol{f}_{sr}$ is exactly the bicubic interpolation; for image denoising, $\boldsymbol{f}_{noise}(I) = I + \eta$, where η is the additive Gaussian noise; for deraining, $\boldsymbol{f}_{rain}(I) = I + r$ in which r is a handcrafted rain streak. The loss function for learning our IPT in the supervised fashion can be formulated as:

$$\mathcal{L}_{supervised} = \sum_{i=1}^{N_t} L_1(\text{IPT}(I^i_{corrupted}), I_{clean}), \tag{4}$$

where L_1 denote the conventional L1 loss for reconstructing desired images and $I^i_{corrupted}$ denote the corrupted image for task i, respectively. In addition, Eq. 4 implies that the proposed framework is trained with multiple image process tasks simultaneously. Specifically, for each batch, we randomly select one task from N_t supervised tasks for training and each task will be processed using the corresponding head, tail and task embedding, simultaneously. After the pre-training the IPT model, it will capture the intrinsic features and transformations for a large variety of image processing tasks thus can be further fine-tuned to apply on the desired task using the new provided dataset. Moreover, other heads and tails will be dropped for saving the computation costs and parameters in the remained head, tail and body will be updated according to the back-propagation.

However, due to the variety of degradation models, we cannot synthesize images for all image processing tasks.

For example, there is a wide range of possible noise levels in practice. Therefore, the generalization ability of the resulting IPT should be further enhanced. Similar to the pre-training natural language processing models, the relationship between patches of images is also informative. The patch in image scenario can be considered as a word in natural language processing. For example, patches cropped from the same feature map are more likely to appear together, which should be embedded into similar positions. Therefore, we introduce contrastive learning [13, 32] for learning universal features so that the pre-trained IPT model can be utilized to unseen tasks. In practice, denote the output patched features generated by IPT decoder for the given input x_j as $f^j_{D_i} \in \mathbb{R}^{P^2 \times C}, i = \{1, \ldots, N\}$, where x_j is selected from a batch of training images $X = \{x_1, x_2, \ldots, x_B\}$. We aims to minimize the distance between patched features from the same images while maximize the distance between patches from different images. The loss function for contrastive learning is formulated as:

$$l(f^j_{D_{i_1}}, f^j_{D_{i_2}}) = -\log \frac{\exp(d(f^j_{D_{i_1}}, f^j_{D_{i_2}}))}{\sum_{k=1}^{B} \mathbb{I}_{k \neq j} \exp(d(f^j_{D_{i_1}}, f^k_{D_{i_2}}))},$$
$$\mathcal{L}_{constrastive} = \frac{1}{BN^2} \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \sum_{j=1}^{B} l(f^j_{D_{i_1}}, f^j_{D_{i_2}}), \tag{5}$$

where $d(a, b) = \frac{a^T b}{\|a\| \|b\|}$ denotes the cosine similarity. Moreover, to make fully usage of both supervised and self-supervised information, we reformulate the loss function as:

$$\mathcal{L}_{IPT} = \lambda \cdot \mathcal{L}_{contrastive} + \mathcal{L}_{supervised}. \tag{6}$$

Wherein, we combine the λ-balanced contrastive loss with the supervised loss as the final objective function of IPT. Thus, the proposed transformer network trained using Eq. 6 can be effectively exploited on various existing image processing tasks.

4. Experiments

In this section, we evaluate the performance of the proposed IPT on various image processing tasks including super-resolution and image denoising. We show that the pre-trained IPT model can achieve state-of-the-art performance on these tasks. Moreover, extensive experiments for ablation study show that the transformer-based models perform better than convolutional neural networks when using the large-scale dataset for solving the image processing problem.

Datasets. To obtain better pre-trained results of the IPT model, we use the well-known ImageNet dataset, which consists of over 1M color images of high diversity. The training images are cropped into 48×48 patches with 3 channels for training, *i.e.*, there are over 10M patches for training the IPT model. We then generate the corrupted images with 6 types of degradation: $2\times, 3\times, 4\times$ bicubic interpolation, $30, 50$ noise level Gaussian noise and adding rain-streaks, respectively. For the rain-streak generation, we follow the method described in [70]. During the test, we crop the images in the test set into 48×48 patches with a 10 pixels overlap. Note that the same testing strategy is also adopted for CNN based models for a fair comparison, and the resulting PSNR values of CNN models are the same as that of their baselines.

Training & Fine-tuning. We use 32 Nvidia NVIDIA Tesla V100 cards to train our IPT model using the conventional Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ for 300 epochs on the modified ImageNet dataset. The initial learning rate is set as $5e^{-5}$ and decayed to $2e^{-5}$ in 200 epoch with 256 batch size. Since the training set consists of different tasks, we cannot input all of them in a single batch due to the expensive memory cost. Therefore, we stack a batch of images from a randomly selected task in each iteration. After pre-training on the entire synthesized dataset, we fine-tune the IPT model on the desired task (*e.g.*, $\times 3$ single image super-resolution) for 30 epochs with a learning rate of $2e^{-5}$. Note that SRCNN [20] also found that using ImageNet training can bring up the performance of the super-resolution task, while we propose a model fitting general low-level vision tasks.

4.1. Super-resolution

We compare our model with several state-of-the-art CNN-based SR methods. As shown in Table 1, our pre-trained IPT outperforms all the other methods and achieves the best performance in $\times 2, \times 3, \times 4$ scale on all datasets. It is worth to highlight that our model achieves 33.76dB PSNR on the $\times 2$ scale Urban100 dataset, which surpasses other methods with more than $\sim$0.4dB, while previous SOTA methods can only achieve a $<$0.2dB improvement compared with others, which indicates the superiority of the proposed model by utilizing large scale pre-training.

We further present the visualization results on our model in $4\times$ scale on Urban100 dataset. As shown in Figure 3, it is difficult for recover the original high resolution images since lots of information are lost due to the high scaling factor. Previous methods generated blurry images, while the super-resolution images produced by our model can well recover the details from the low-resolution images.

4.2. Denoising

Since our pre-trained model can be well adapt to many tasks, we then evaluate the performance of our model on image denoising task. The training and testing data is generated by adding Gaussian noise with $\sigma = 30, 50$ to the clean images.

To verify the effectiveness of the proposed method,

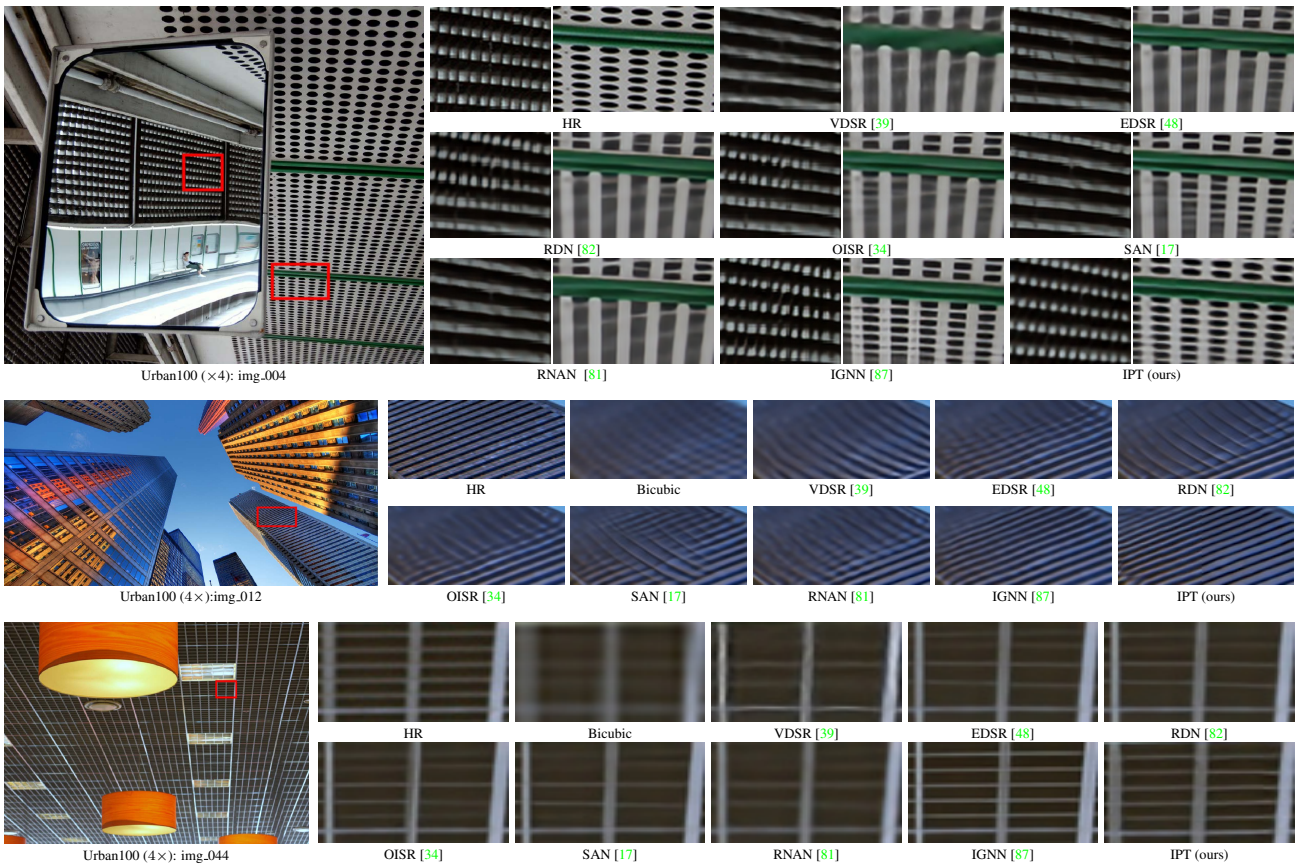

Figure 3. Visual results with bicubic downsampling (×4) from Urban100. The proposed method recovers more details. Compared images are derived from [87].

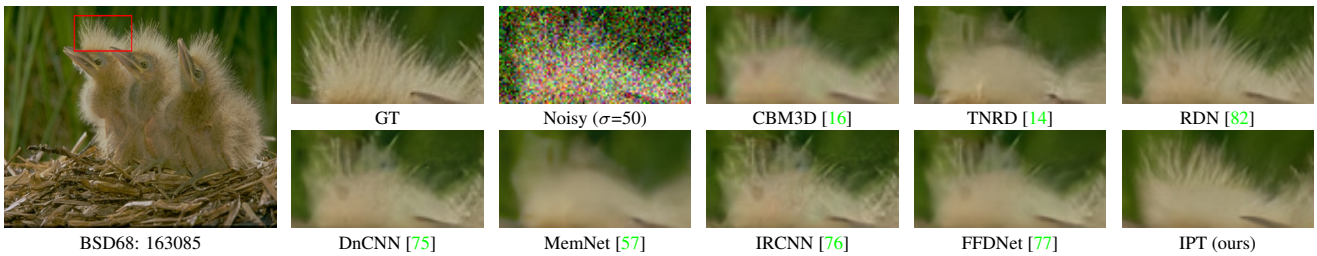

Figure 4. Color image denoising results with noise level $\sigma = 50$. Compared images are derived from [78].

we compare our results with various state-of-the-art models. Table 2 reported the color image denoising results on BSD68 and Urban100 dataset. As a result, our IPT achieves the best results among all denoising methods on different Gaussian noise level. Moreover, we surprisingly found that our model improve the state-of-the-art performance by ∼2dB on the Urban100 dataset, which demonstrate the effectiveness of pre-training and the superiority of our transformer-based model.

Figure 4 shows the visualization of the resulted images. As shown in the figure, noisy images are hard to be recognized and it is difficult to recover the clean images. Therefore, existing methods fail to reconstruct enough details and generate abnormal pixels. As a result, our pre-trained model can well recover several details in the hair of this cat and our visual quality beats all the previous models obviously.

4.3. Deraining

For the image deraining task, we evaluate our model on the synthesized Rain100L dataset [70], which consists of 100 rainy images. Quantitative results can be viewed in Table 3. Compared with the state-of-the-art methods, we achieve the best performance (41.62dB) with an 1.62dB improvement.

Figure 5 shows the visualization results. Previous methods are failed to reconstruct the original clean images since they lack of image prior. As a result, our IPT model can present exactly the same image as the ground-truth and sur-

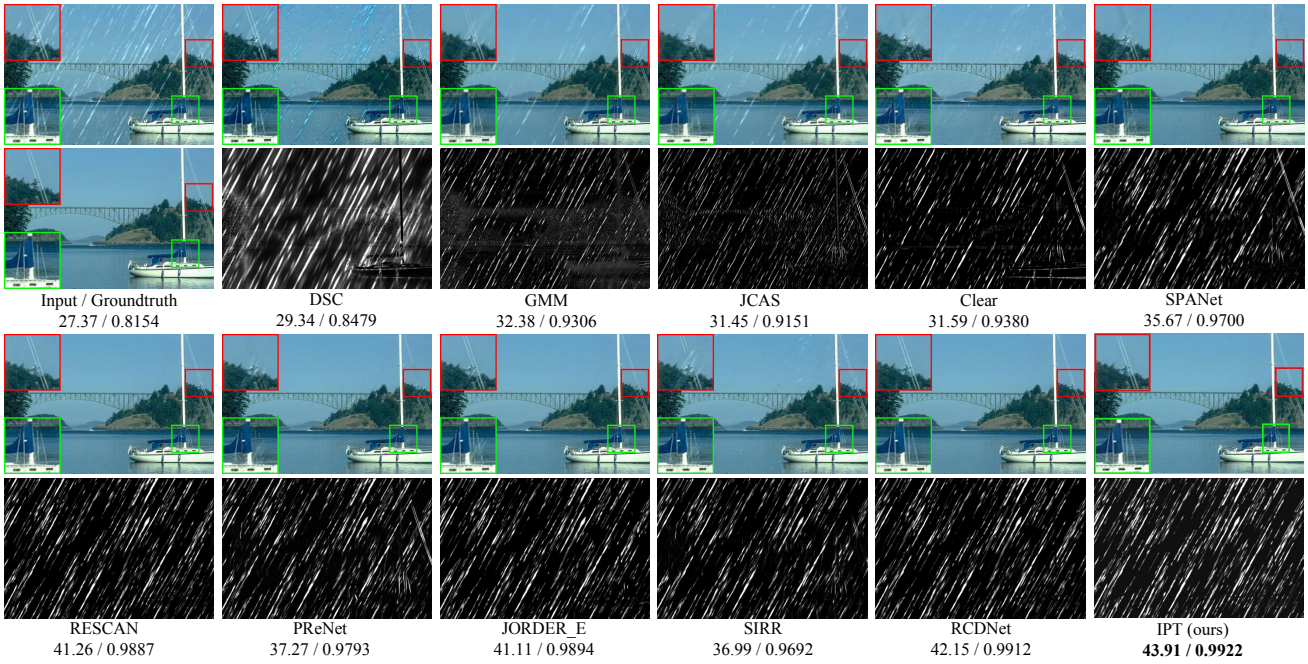

Figure 5. Image deraining results on the Rain100L dataset. Compared images are derived from [63].

Table 1. Quantitative results on image super-resolution. Best and second best results are **highlighted** and underlined.

Method	Scale	Set5	Set14	B100	Urban100
VDSR [39]	×2	37.53	33.05	31.90	30.77
EDSR [48]	×2	38.11	33.92	32.32	32.93
RCAN [80]	×2	38.27	34.12	32.41	33.34
RDN [82]	×2	38.24	34.01	32.34	32.89
OISR-RK3 [34]	×2	38.21	33.94	32.36	33.03
RNAN [81]	×2	38.17	33.87	32.32	32.73
SAN [17]	×2	38.31	34.07	32.42	33.10
HAN [51]	×2	38.27	34.16	32.41	33.35
IGNN [87]	×2	38.24	34.07	32.41	33.23
IPT (ours)	×2	**38.37**	**34.43**	**32.48**	**33.76**
VDSR [39]	×3	33.67	29.78	28.83	27.14
EDSR [48]	×3	34.65	30.52	29.25	28.80
RCAN [80]	×3	34.74	30.65	29.32	29.09
RDN [82]	×3	34.71	30.57	29.26	28.80
OISR-RK3 [34]	×3	34.72	30.57	29.29	28.95
RNAN [81]	×3	34.66	30.52	29.26	28.75
SAN [17]	×3	34.75	30.59	29.33	28.93
HAN [51]	×3	34.75	30.67	29.32	29.10
IGNN [87]	×3	34.72	30.66	29.31	29.03
IPT (ours)	×3	**34.81**	**30.85**	**29.38**	**29.49**
VDSR [39]	×4	31.35	28.02	27.29	25.18
EDSR [48]	×4	32.46	28.80	27.71	26.64
RCAN [80]	×4	32.63	28.87	27.77	26.82
SAN [17]	×4	**32.64**	28.92	27.78	26.79
RDN [82]	×4	32.47	28.81	27.72	26.61
OISR-RK3 [34]	×4	32.53	28.86	27.75	26.79
RNAN [81]	×4	32.49	28.83	27.72	26.61
HAN [51]	×4	**32.64**	28.90	27.80	26.85
IGNN [87]	×4	32.57	28.85	27.77	26.84
IPT (ours)	×4	**32.64**	**29.01**	**27.82**	**27.26**

passes all the previous algorithms in visual quality. This result substantiates the generality of the proposed model.

Table 2. Quantitative results on color image denoising. Best and second best results are **highlighted** and underlined.

Method	BSD68		Urban100	
	30	50	30	50
CBM3D [16]	29.73	27.38	30.36	27.94
TNRD [14]	27.64	25.96	27.40	25.52
DnCNN [75]	30.40	28.01	30.28	28.16
MemNet [57]	28.39	26.33	28.93	26.53
IRCNN [76]	30.22	27.86	30.28	27.69
FFDNet [77]	30.31	27.96	30.53	28.05
SADNet [9]	30.64	28.32	N/A	N/A
RDN [83]	30.67	28.31	31.69	29.29
IPT (ours)	**32.32**	**29.88**	**33.75**	**31.12**

4.4. Generalization Ability

Although we can generate various corrupted images, natural images are of high complexity and we cannot synthesize all possible images for pre-training the transformer model. However, a good pre-trained model should have the capacity for well adapting other tasks as those in the field of NLP. To this end, we then conduct several experiments to verify the generalization ability of our model. In practice, we test corrupted images that did not include in our synthesized ImageNet dataset, *i.e.*, image denoising with noisy level 10 and 70, respectively. We use the heads and tails for image denoising tasks as the pre-trained model.

The detailed results are shown in Table 4, we compare the performance of using the pre-trained IPT model and the state-of-the-art methods for image denoising. Obviously, IPT model outperforms other conventional methods, which

Table 3. Quantitative results of image deraining on the Rain100L dataset. Best and second best results are **highlighted** and underlined.

Method	Input	DSC [28]	GMM [46]	JCAS [30]	Clear [27]	DDN [28]
PSNR	26.90	27.34	29.05	28.54	30.24	32.38
SSIM	0.8384	0.8494	0.8717	0.8524	0.9344	0.9258
RESCAN [45]	PReNet [55]	JORDER_E [70]	SPANet [65]	SSIR [67]	RCDNet [63]	IPT (ours)
38.52	37.45	38.59	35.33	32.37	40.00	**41.62**
0.9812	0.9790	0.9834	0.9694	0.9258	0.9860	**0.9880**

Table 4. Generation ability of our IPT model on color image denoising with different noise levels. Best and second best results are **highlighted** and underlined.

Method	BSD68		Urban100	
	10	70	10	70
CBM3D [16]	35.91	26.00	36.00	26.31
TNRD [14]	33.36	23.83	33.60	22.63
DnCNN [75]	36.31	26.56	36.21	26.17
MemNet [57]	N/A	25.08	N/A	24.96
IRCNN [76]	36.06	N/A	35.81	N/A
FFDNet [77]	36.14	26.53	35.77	26.39
RDN [83]	36.47	26.85	36.69	27.63
IPT (ours)	**38.30**	**28.21**	**39.07**	**28.80**

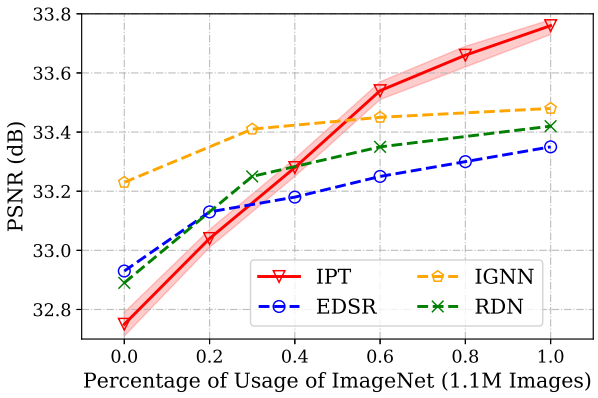

Figure 6. The performance of CNN and IPT models using different percentages of data.

demonstrates that the pre-trained model can capture more useful information and features from the large-scale dataset.

4.5. Ablation Study

Impact of data percentage. To evaluate the effectiveness of the transformer architecture, we conduct experiments to analyse the improvement of pre-training on CNN-based model and transformer-based model. We use 20%, 40%, 60%, 80% and 100% percentages of the synthesized ImageNet dataset to analyse the impact on the number of used data for resulting performance. Figure 6 shows the results of different pre-trained models. When the models are not pre-trained or pre-trained with small amount ($<$ 60%) of the entire dataset, the CNN models achieve better performance. In contrast, when using large-scale data, the transformer-based models overwhelming CNN models, which demonstrates that the effectiveness of our IPT model for pre-training.

Table 5. Impact of λ for contrastive learning.

λ	0	0.05	0.1	0.2	0.5
PSNR	38.27	38.32	38.37	38.33	38.26

Impact of contrastive learning. As discussed above, to improve the representation ability of our pre-trained model, we embed the contrastive learning loss (Eq. 6) into the training procedure. We then evaluate its effectiveness on the $\times 2$ scale super-resolution task using the Set4 dataset. Table 5 shows the impact of the hyper-parameter λ for balancing the two terms in Eq. 6. When λ=0, the IPT model is trained using only a supervised learning approach, the resulting PSNR value is 38.27dB. When employing the contrastive loss for self-supervised learning, the model can achieve a 38.37dB PSNR value ($\lambda = 0.1$), which is about 0.1dB higher than that of the model trained with $\lambda = 0$. These results further demonstrate the effectiveness of the contrastive learning for learning better pre-trained IPT model.

5. Conclusions and Discussions

This paper aims to address the image processing problems using a pre-trained transformer model (IPT). The IPT model is designed with multi-heads,multi-tails a shared transformer body for serving different image processing task such as image super-resolution and denoising. To maximally excavate the performance of the transformer architecture on various tasks, we explore a synthesized ImageNet datesets. Wherein, each original image will be degraded to a series of counterparts as paired training data. The IPT model is then trained using supervised and self-supervised approaches which shows strong ability for capturing intrinsic features for low-level image processing. Experimental results demonstrate that our IPT can outperform the state-of-the-art methods using only one pre-trained model after a quickly fine-tuning. In the future work, we will extend our IPT model to more tasks such as inpainting, dehazing, *etc*.

Acknowledgment This work is supported by National Natural Science Foundation of China under Grant No. 61876007, and Australian Research Council under Project DE180101438 and DP210101859.

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. 4

[2] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 252–268, 2018. 2

[3] Saeed Anwar and Nick Barnes. Densely residual laplacian super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[5] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[6] 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

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020. 2, 3

[8] Gabriella Castellano, Leonardo Bonilha, LM Li, and Fernando Cendes. Texture analysis of medical images. *Clinical radiology*, 59(12):1061–1069, 2004. 1

[9] Meng Chang, Qi Li, Huajun Feng, and Zhihai Xu. Spatial-adaptive network for single image denoising. *arXiv preprint arXiv:2001.10291*, 2020. 7

[10] Liang Chen, Faming Fang, Shen Lei, Fang Li, and Guixu Zhang. Enhanced sparse model for blind deblurring. 2020. 2

[11] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 1, 2020. 3

[12] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Michael Carbin, and Zhangyang Wang. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. *arXiv preprint arXiv:2012.06908*, 2020. 1

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 5

[14] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1256–1272, 2016. 6, 7, 8

[15] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2019. 3

[16] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Color image denoising via sparse 3d collaborative filtering with grouping constraint in luminance-chrominance space. In *2007 IEEE International Conference on Image Processing*, volume 1, pages I–313. IEEE, 2007. 6, 7, 8

[17] 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 conference on computer vision and pattern recognition*, pages 11065–11074, 2019. 6, 7

[18] 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. 1

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[20] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 2, 5

[21] 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. 2

[22] 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. 2, 3

[23] Thomas Eboli, Jian Sun, and Jean Ponce. End-to-end interpretable learning of non-blind image deblurring. *arXiv preprint arXiv:2007.01769*, 2020. 2

[24] Yuchen Fan, Honghui Shi, Jiahui Yu, Ding Liu, Wei Han, Haichao Yu, Zhangyang Wang, Xinchao Wang, and Thomas S Huang. Balanced two-stage residual networks for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 161–168, 2017. 2

[25] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 2

[26] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 3

[27] Xueyang Fu, Jiabin Huang, Xinghao Ding, Yinghao Liao, and John Paisley. Clearing the skies: A deep network archi-

tecture for single-image rain removal. *IEEE Transactions on Image Processing*, 26(6):2944–2956, 2017. 8

[28] 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. 8

[29] Xueyang Fu, Borong Liang, Yue Huang, Xinghao Ding, and John Paisley. Lightweight pyramid networks for image deraining. *IEEE transactions on neural networks and learning systems*, 2019. 2

[30] Shuhang Gu, Deyu Meng, Wangmeng Zuo, and Lei Zhang. Joint convolutional analysis and synthesis sparse representation for single image layer separation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1708–1716, 2017. 4, 8

[31] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1712–1722, 2019. 2

[32] 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. 5

[33] 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. 2

[34] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. Ode-inspired network design for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1732–1741, 2019. 6, 7

[35] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8022–8031, 2019. 2

[36] Xixi Jia, Sanyang Liu, Xiangchu Feng, and Lei Zhang. Focnet: A fractional optimal control network for image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6054–6063, 2019. 2

[37] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2070–2079, 2019. 3

[38] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong gan. *arXiv preprint arXiv:2102.07074*, 2021. 3

[39] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 2, 6, 7

[40] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6, 2019. 3

[41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 2

[42] Stamatios Lefkimmiatis. Non-local color image denoising with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3587–3596, 2017. 2

[43] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. An all-in-one network for dehazing and beyond. *arXiv preprint arXiv:1707.06543*, 2017. 2

[44] Siyuan Li, Wenqi Ren, Feng Wang, Iago Breno Araujo, Eric K Tokuda, Roberto Hirata Junior, Roberto M Cesar-Jr, Zhangyang Wang, and Xiaochun Cao. A comprehensive benchmark analysis of single image deraining: Current challenges and future perspectives. *International Journal of Computer Vision*, pages 1–22, 2021. 2

[45] 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. 8

[46] 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. 8

[47] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2

[48] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 6, 7

[49] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[50] Boyu Lu, Jun-Cheng Chen, and Rama Chellappa. Unsupervised domain-specific deblurring via disentangled representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10225–10234, 2019. 2

[51] 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. 7

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

[53] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[54] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 2

[55] 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 conference on computer vision and pattern recognition*, pages 3937–3946, 2019. 2, 8

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[57] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pages 4539–4547, 2017. 6, 7, 8

[58] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[59] 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

[60] Chunwei Tian, Yong Xu, Zuoyong Li, Wangmeng Zuo, Lunke Fei, and Hong Liu. Attention-guided cnn for image denoising. *Neural Networks*, 124:117–129, 2020. 2

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 4

[62] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 2

[63] Hong Wang, Qi Xie, Qian Zhao, and Deyu Meng. A model-driven deep neural network for single image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3103–3112, 2020. 7, 8

[64] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*, 2019. 2

[65] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12270–12279, 2019. 2, 8

[66] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3

[67] Wei Wei, Deyu Meng, Qian Zhao, Zongben Xu, and Ying Wu. Semi-supervised transfer learning for image rain removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2019. 8

[68] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020. 3

[69] Wenhan Yang, Jiaying Liu, Shuai Yang, and Zongming Guo. Scale-free single image deraining via visibility-enhanced recurrent wavelet learning. *IEEE Transactions on Image Processing*, 28(6):2948–2961, 2019. 2

[70] Wenhan Yang, Robby T Tan, Jiashi Feng, Zongming Guo, Shuicheng Yan, and Jiaying Liu. Joint rain detection and removal from a single image with contextualized deep networks. *IEEE transactions on pattern analysis and machine intelligence*, 42(6):1377–1393, 2019. 5, 6, 8

[71] Xitong Yang, Zheng Xu, and Jiebo Luo. Towards perceptual image dehazing by physics-based disentanglement and adversarial training. In *AAAI*, pages 7485–7492, 2018. 2

[72] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018. 3

[73] Yongnian Zeng, Wei Huang, Maoguo Liu, Honghui Zhang, and Bin Zou. Fusion of satellite images in urban area: Assessing the quality of resulting images. In *2010 18th International Conference on Geoinformatics*, pages 1–4. IEEE, 2010. 1

[74] He Zhang and Vishal M Patel. Densely connected pyramid dehazing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3194–3203, 2018. 2

[75] 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. 6, 7, 8

[76] 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. 6, 7, 8

[77] 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. 6, 7, 8

[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. 6

[79] Songyang Zhang, Xuming He, and Shipeng Yan. Latentgnn: Learning efficient non-local relations for visual recognition. In *International Conference on Machine Learning*, pages 7374–7383, 2019. 3

[80] 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. 2, 7

[81] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 6, 7

[82] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 6, 7

[83] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 7, 8

[84] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10076–10085, 2020. 3

[85] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for rgbd salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3927–3936, 2019. 1

[86] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8779–8788, 2019. 1

[87] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. *Advances in Neural Information Processing Systems*, 33, 2020. 6, 7

[88] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3