Pre-Trained Image Processing Transformer [1]

Minji Kim (minji@snu.ac.kr; 2020-28702), Dept. of Electrical and Computer Engineering, Seoul National University

1. Introduction

This paper presents a pre-trained transformer model for low-level computer vision tasks, *e.g.*, denoising, deraining, and super-resolution. For pre-training in a large-scale benchmark, the ImageNet is adopted for generating a large amount of image pairs. The proposed model is end-to-end trained on these pairs with multi-heads and multi-tails for various low-level vision tasks. After fine-tuning, IPT outperforms the current state-of-the-art models as shown in Fig. 1.

2. Image Processing Transformer (IPT)

2.1. Architecture

Fig. 2 shows the overall architecture of IPT. The encoder and decoder architecture basically follows the original Transformer [3]. The multi-head and multi-tail is constructed for different tasks, *i.e.*, denoising, deraining, and super-resolution for 2× and 4× scaling. The input images are first embedded into visual features and then divided into patches as similar to Vision Transformer [2]. The resulting images are reconstructed by ensembling output patches.

2.2. ImageNet Pre-training

The challenge of pre-training a low-level vision task model is the lack of large-scale dataset. To tackle this issue, the authors generated the synthesized images from ImageNet based on a degradation transformation for each task, *e.g.*, bicubic interpolation for the super-resolution task, additive Gaussian noise for the denoising task. Finally, the IPT model is trained to reconstruct the clean image from the corrupted one. Additionally, the contrastive loss is adopted for minimizing the distance between features from same images while maximizing the distance between features from different images.

3. Personal Notes

The IPT model is a unified pre-trained Transformer model for several image processing tasks. Although this model achieves state-of-the-art performance, I was not able to find any other novelty or contribution. Besides, this paper is not well-written, and thus it was hard for me to catch the motivation of the method.

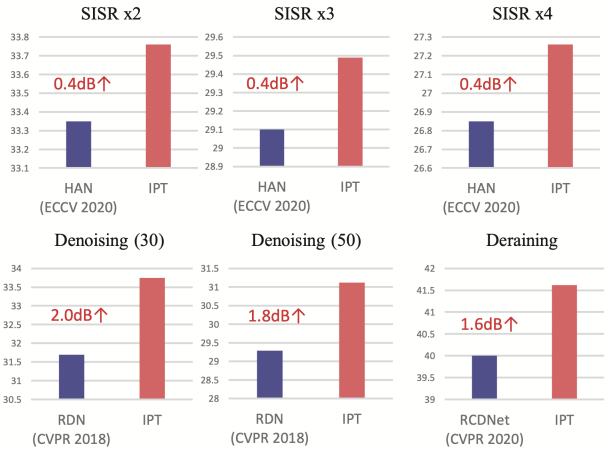

Figure 1. Comparison on the performance of IPT and state-of-the-art image processing models on different tasks.

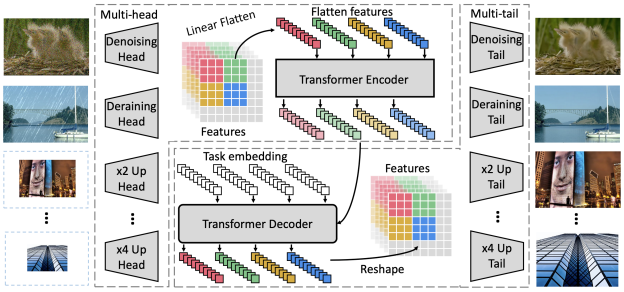

Figure 2. The overall pipeline of IPT. IPT consists of multi-head and multi-tail for different tasks and a shared Transformer body including encoder and decoder.

References

[1] H. Chen, et al. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 1

[2] A. Dosovitskiy, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[3] A. Vaswani, et al. Attention is all you need. In *NeurIPS*, 2017. 1