Image Denoising using Swin-Transformer UNet

Kai San Chan, Huimiao Chen, Jie Song, Sile Wang

May 14, 2023

1 Abstract

Image restoration, which involves the process of recovering degraded images by removing noise, is significantly improved through image denoising. This is a critical technique inspired by the pioneering work of Fan [10] and Liang [17]. We've formulated a model named Shift-Transformer UNet (SUNet). SUNet integrates the foundations of UNet with the Shift-window Transformer (Swin-Transformer). This is accomplished by substituting all convolutional blocks with shift window multihead attention layers from the Shift-window Transformer. We trained the model using open-source datasets DIV2K and Flickr2K, and tested it on BSD100, BSD200, and General100. SUNet outperforms both UNet and Swin-Transformer, as evidenced by its superior results in Peak Source to Noise Ratio and the Structural Similarity Index Measure. Interestingly, SUNet demonstrates enhanced capabilities compared to UNet. While the latter can only recover the structure of an object from a noisy image, SUNet can reconstruct detailed information that is often overlooked by UNet, thanks to the incorporation of Swin-Transformer. Future research should delve deeper into the potential of the Swin Transformer.

2 Introduction

Image denoising is an integral aspect of image restoration. Its primary function is to estimate and recover a clean image from a noisy, degraded, or blurred version [28]. Noise often originates during transmission and compression processes from an unknown latent observation [36]. Hence, the significance of image denoising techniques in extracting noise and retrieving the latent observation from the provided noisy image cannot be overstated [29]. Additionally, denoising techniques find a broad spectrum of applications in diverse fields such as film restoration, medical imaging processing, surveillance, and autopilot [28, 15].

Over the past half-century, image denoising has been the subject of considerable attention and interest [2, 37]. As a result, a host of competitive image denoising methods have been proposed, including the Markov random field (MRF) [25], weighted nuclear norm minimization (WNNM) [12], and trainable nonlinear reaction diffusion (TNRD) [7], among others. Despite the relatively satisfactory performance of these methods, they are hampered by the need for manual parameter setting, simplified tasks, and test phase optimization methodologies [19].

The evolution of deep convolutional neural networks has significantly influenced image denoising and restoration, resulting in enhanced performance [23, 30]. Moreover, the rise of Transformer-based models in recent years [33, 16, 26] has led to remarkable outcomes in various vision tasks, such as recognition [9, 5, 13]. This growth has also facilitated the application of the attention mechanism to image restoration [6, 17, 35]. Consequently, the advances in both the attention mechanism and deep neural networks have paved the way for designing a comprehensive denoising model that effectively combines the benefits of both technologies.

3 Related Work

Contemporary state-of-the-art denoising or restoration methods largely employ auto-encoder-based architectures like Real-ESRGAN and DMPHN [8, 21, 34, 39]. The advantageous features of these models, such as end-to-end training, skip connection, and multi-level learning, can be attributed to the auto-encoder, especially UNet. Concurrently, Transformer-based architectures have exhibited superior performance in natural language tasks [3, 11, 22], while their attention mechanism has proven advantageous in the computer vision domain [9, 32]. As a result, an increasing number of studies on image denoising tasks have leveraged the benefits of auto-encoder-based architectures and the attention mechanism by designing combinations of auto-encoders and Transformers, such as Uformer [35] and Restormer [38].

Among all transformer-based models, the Shifted windows-Transformer (Swin-Transformer) shows the most promise to replace the backbone of traditional convolutional blocks. By limiting self-attention computation to non-overlapping local windows and allowing for cross-window connections, it further reduces computational complexity and enhances efficiency [18]. Taking inspiration from Fan [10] and Liang [17], we aim to merge the Auto-Encoder Unet with the Swin-Transformer for our denoising task. This resulted in the proposal of the Swin-Transformer UNet (SUNet). SUNet capitalizes on the skip connection feature of the traditional UNet, and is further enhanced by the Swin-Transformer backbone, which increases efficiency and enables the model to better focus on detail by tracking the correlation between each pixel within the windows.

4 Method

The general structure of the Swin-Transformer UNet (SUNet) architecture, grounded in the principles of UNet [24] and Swin-Transformer [18], is illustrated in Figure 1. SUNet encompasses three primary components: 1) Shallow feature extraction; 2) UNet feature extraction; and 3) Reconstruction module. Initially, the input image is processed through the shallow feature extraction module to obtain a low-level noisy feature map. This noisy feature map is then fed into the UNet feature extraction module to generate a reconstructed feature map. Finally, the Reconstruction module employs this reconstructed feature map to recover the final image. Each module is further elaborated in the subsequent subsections.

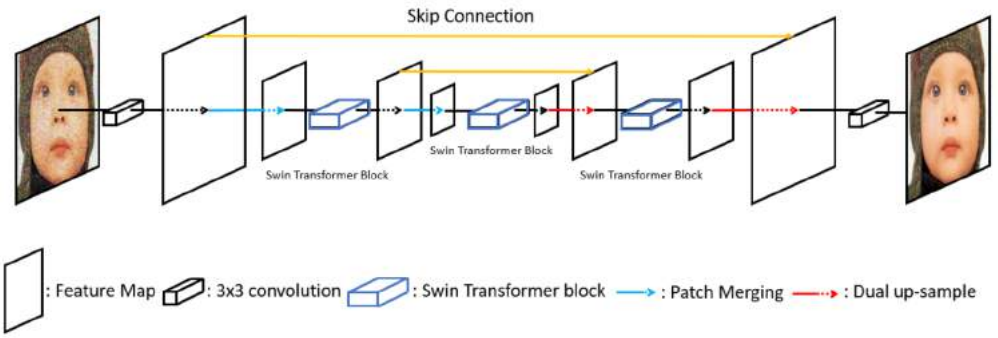


Figure 1: Structure of Swin-Transformer UNet (SUNet)

For a given noisy input image $X \in R^{H \times W \times 3}$, where H and W represent the image resolution, and 3 corresponds to the RGB channels, we utilize a single 3×3 convolution layer $M_{SFE}(\cdot)$ for feature extraction. After empirical testing, the output channels from the Shallow Feature Extraction are set to 96. The extracted feature can be defined as:

$$F_{SFE} = M_{SFE}(X) \tag{1}$$

Next, the shallow feature F_{SFE} is input into the UNet Feature Extraction module $M_{UFE}(\cdot)$, thereby

extracting multi-level and multi-scale deep features. UNet Feature Extraction can be expressed as:

$$F_{UFE} = M_{UFE}(F_{SFE}) \tag{2}$$

The module M_{UFE} is where the integration of the Swin-Transformer and UNet occurs, with all convolutional layers being replaced by Swin-Transformer blocks. Each Swin-Transformer block incorporates 8 Swin Transformer Layers. Detailed descriptions of the Swin-Transformer Block and Swin-Transformer Layer are provided below.

Figure 2 illustrates the structure of the Swin-Transformer block (on the left) and Swin-Transformer Layer (on the right) that replace the traditional convolutional layer. As shown on the left, each Swin-Transformer Block (STB) comprises 8 Swin-Transformer Layers (STLs). On the right, the architecture of the Swin-Transformer Layer is detailed. This layer includes the Normalization Layer (Layer Norm), Window Multi-Head Self-Attention module (Window MSA), Shift-window Multi-Head Self-Attention (Shift Window MSA), Multi-Layer Perceptron (MLP), and residual connections.

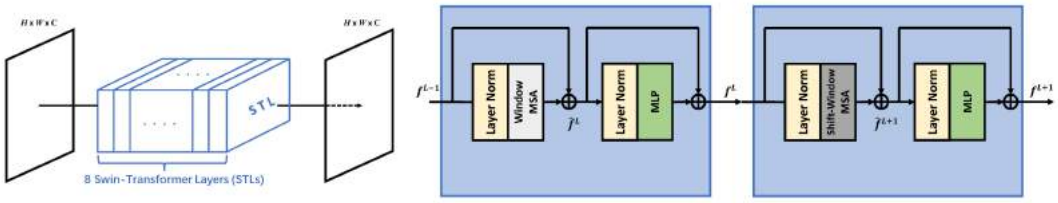


Figure 2: Swin-Transformer Block (STB) is shown on the left, Swin-Transformer Layer is shown on the right

The design of the STL is inspired by the original Transformer layer from the field of Natural Language Processing (NLP). In this context, the STL is divided into two sections, namely, window multi-head self-attention (Window-MSA) and shifted-Window Multi-Head self-attention (Shift-Window MSA). The application of the cyclic shift technique enhances computational efficiency and preserves the inherent characteristics of convolution, including translation invariance, rotation invariance, and the size-independent relationship between the receptive field and layers. Such properties allow it to replace the convolutional block without losing the model's invariance characteristics. Consequently, the Swin-Transformer Block can regulate the resolution (H, W) and the number of channels (C) in the output features, akin to the convolution operation. The detailed function can be expressed as:

$$\hat{f}^L = Window - MSA(LN(f^{L-1})) + f^{L-1} \tag{3}$$

$$f^L = MLP(LN(\hat{f}^L)) + \hat{f}^L \tag{4}$$

$$\hat{f}^{L+1} = ShiftWindow - MSA(LN(f^L)) + f^L \tag{5}$$

$$f^{L+1} = MLP(LN(f \hat{+} 1^L)) + f \hat{+} 1^L \tag{6}$$

where $LN(\cdot)$ represents Layer Normalization and $MLP(\cdot)$ signifies a multilayer perceptron consisting of two fully connected layers with the Gaussian Error Linear Unit (GELU) as the activation function.

In the case of the down-sampling module, we follow the strategy proposed by Liu et al. [18] and Cao et al. [4], where 2×2 neighboring patches of input features are concatenated, followed by the use of a linear layer to obtain the desired output channel number. This can be seen as the initial step of convolution. For the up-sampling path, we make use of the Dual up-sample module, which is shown to be an effective counterpart to transpose convolution in the up-sampling module [18, 4, 10]. This module integrates two prevalent up-sampling methods (i.e., Bilinear and PixelShuffle [27]) in a bid to

avert checkerboard artifacts. The architecture of the suggested up-sampling module is depicted in the figure below.

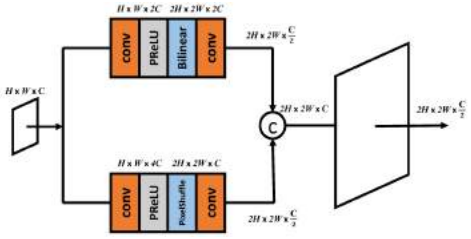


Figure 3: Dual Up-sampling module

5 Experiment

For evaluating the SUNet, we draw upon the original UNet and Swin-Transformer for model comparison. All these models are end-to-end trainable without using any pre-trained networks.

The datasets used for training include DIV2K (900 images) [1] and Flickr2K (2650 images) [31], summing up to 3550 images in total. The testing dataset comprises BSD100 (100 images) [20], BSD200 (200 images), and General100 (100 images) [14], totalling 400 images. The input images are corrupted with Gaussian noise with zero mean and a random standard deviation within the range [0,0.8]. The training set is further divided into training and validation subsets, leading to a final ratio of around 6:1:1 (3018:532:400) for training, validation, and testing sets.

We use Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) as the metrics for quantitative evaluation. Both PSNR and SSIM values are higher-the-better, with PSNR unit being decibel (dB) and SSIM ranging from zero to one. The PSNR and SSIM are calculated as follows:

$$PSNR(x, y) = 10 \log_{10} \left(\frac{\max(\max(x), \max(y))}{\|x - y\|_2^2} \right) \tag{7}$$

$$SSIM(x, y) = \frac{(2\mu_x \mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c1)(\sigma_x^2 + \sigma_y^2 + c2)} \tag{8}$$

where x is the output and y is the label image.

All models use L2 Loss function:

$$L2Loss(x, y) = \|x - y\|_2^2 \tag{9}$$

where x is the output and y is the label. We use Adam optimizer with a batch size of 32 and 1e-4 initial learning rate. We set the total training epochs to 600 with early stopping. The model weights with the lowest validation loss are saved and used on the test dataset for model comparison. Besides the quantitative results on the test set in terms of PSNR and SSIM, we also present several examples to further investigate the performance.

6 Results and Analysis

The quantitative results showcased in the Table reveal that SUNet outperforms both UNet and Swin-Transformer in terms of PSNR and SSIM. Specifically, SUNet demonstrates superior performance, surpassing the second-best model, UNet, with a noticeable margin. This significant improvement is observed in both the PSNR and SSIM metrics, indicating the effectiveness of the proposed SUNet model in image denoising tasks.

Table 1: Model Comparison

	PSNR	SSIM
Input	25.64	0.2713
UNet [24]	30.72	0.6680
Swin-Transformer [17]	29.04	0.6066
SUNet [10]	**30.93**	**0.6757**

As illustrated in the sample output from the BSD200 test set, all of the models successfully restore the majority of the noisy image, demonstrating their practical utility in image denoising tasks. Notably, SUNet offers a superior performance in terms of PSNR, suggesting that it is able to more accurately restore image details from noise. On the other hand, UNet stands out in terms of SSIM, implying its efficiency in maintaining the structural similarity of the restored image to the original one.

Ground Truth Label

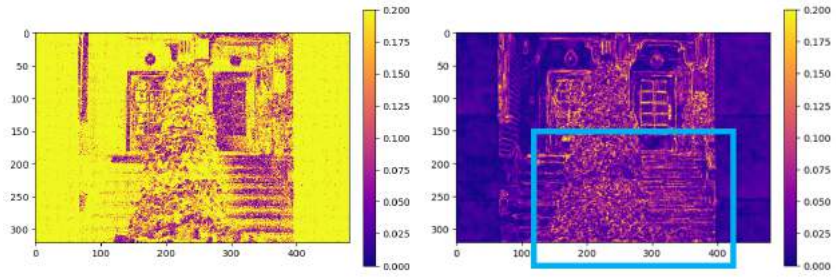
Input noisy image
PSNR: 14.49 SSIM: 0.2357

SUNet
PSNR: 47.25 SSIM: 0.7104

UNet
PSNR: 42.21 **SSIM: 0.7449**

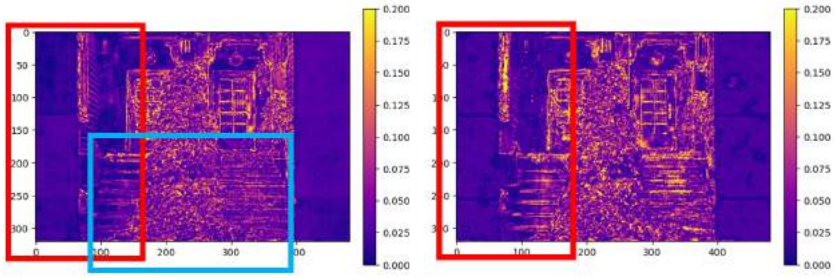
SWINIR
PSNR: 40.39 SSIM: 0.6559

To gain a deeper understanding and clearer view of the reconstruction process, we generate the absolute difference maps between the output and label images, noting that a lower value corresponds to a superior recovery. Upon comparison, it becomes evident that SUNet manages to recover far more details than UNet, as highlighted in the blue box. Additionally, compared to Swin-Transformer, SUNet demonstrates superior structural reconstruction, as illustrated in the red box. This suggests that SUNet can not only recover the structure of the objects from a noisy image akin to UNet, but also reconstruct detailed information that was missed in the original UNet. This explains why SUNet exhibits similar, albeit slightly higher overall PSNR and SSIM compared to UNet. The enhancements in SUNet are attributed to the superior recovery of detailed information. While these minute details might not significantly elevate the overall quantitative results, they make a substantial difference in terms of perceptual quality.



Noisy image
PSNR: 14.49 SSIM: 0.2357

UNet
PSNR: 42.21 **SSIM: 0.7449**



SUNet
PSNR: 47.25 SSIM: 0.7104

SWINIR
PSNR: 40.39 SSIM: 0.6559

The figure below presents another illustrative example. In this case, SUNet exhibits superior performance with respect to SSIM, while Swin-Transformer leads in terms of PSNR. This instance further substantiates the observation that UNet excels in reconstructing the structure of objects, whereas Swin-Transformer is more adept at recovering finer details. The results and examples corroborate that SUNet indeed leverages the strengths of both UNet and Swin-Transformer.

Ground Truth Label

Input noisy image
PSNR: 18.54, SSIM: 0.3094

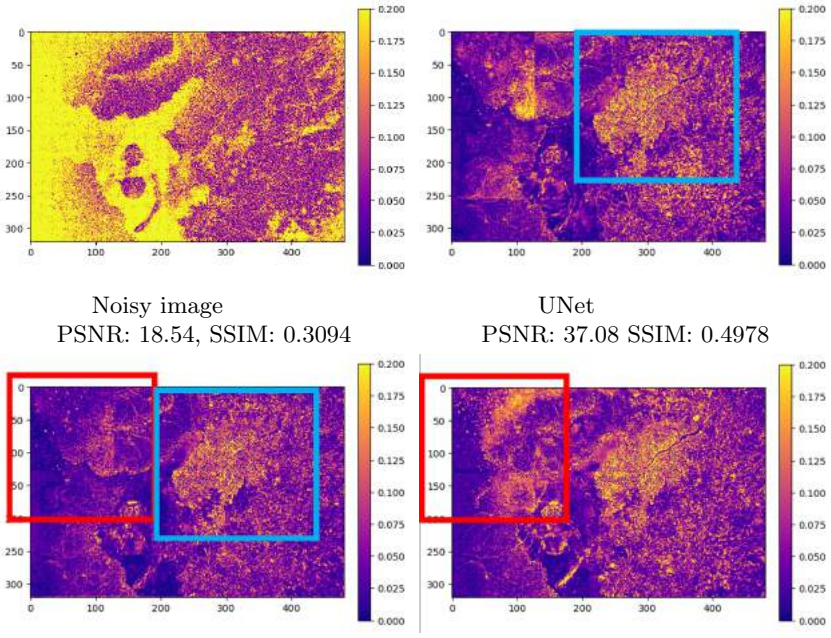
SUNet
PSNR: 37.71 **SSIM: 0.5358**

UNet
PSNR: 37.08 SSIM: 0.4978

SWINIR
PSNR: 37.80 SSIM: 0.5105

The absolute difference map is also plotted to assess performance. It is evident that this image contains a profusion of details and structures, which explains why all models perform less optimally as compared to the previous image. Further research and studies can be undertaken to enhance the performance on intricate images such as this one. A comparison of the blue boxes for UNet and SUNet reveals that SUNet recovers more detailed information. Similarly, comparing the red boxes for Swin-Transformer

and SUNet, it is apparent that more structure is reconstructed in the top left region by SUNet. These observations are in line with our findings and the objectives that guided the design of our model.



Noisy image
PSNR: 18.54, SSIM: 0.3094

UNet
PSNR: 37.08 SSIM: 0.4978

SUNet
PSNR: 37.71 **SSIM: 0.5358**

SWINIR
PSNR: 37.80 SSIM: 0.5105

7 Conclusion

In this work, we present a deep learning model, SUNet, for the task of image denoising, which cleverly amalgamates the UNet and Shift-window Transformer by replacing all convolutional blocks in the auto-encoder UNet with shift-window multi-head self-attention layers. This model also incorporates dual upsampling blocks to facilitate superior upsample merging. Experiments are conducted on several renowned open datasets, namely DIV2K, Flickr2K, BSD100, BSD200, and General100. The results demonstrate that SUNet outperforms both UNet and Swin-Transformer, achieving superior PSNR and SSIM values. A closer inspection of the reconstructed images and the model architecture reveals that SUNet possesses the ability to restore not only the structural integrity of noisy images akin to UNet, but also the detailed information often overlooked by UNet, thanks to the integration of the Swin-Transformer. However, it may be premature to assert that the Swin Transformer can entirely supersede convolution. Future research should further explore the potential of Swin Transformer. Our future endeavours will attempt to tackle more complex denoising or restoration tasks, such as those involving real-world noise and blur, while continuing to employ Swin-Transformer Layers.

8 Acknowledgement

We extend our special acknowledgement to Fan [10] and Liang [17], the original designers of SUNet and SwinIR, respectively. All modifications were made based on the official SUNet codebase. The official GitHub repositories for SUNet and SwinIR can be found at:

`https://github.com/FanChiMao/SUNet`
`https://github.com/JingyunLiang/SwinIR`

9 More Result

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[2] Reinhard Bernstein. Adaptive nonlinear filters for simultaneous removal of different kinds of noise in images. *IEEE Transactions on Circuits and Systems*, 34(11):1275–1291, 1987.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 205–218. Springer, 2023.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

[7] 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.

[8] Shen Cheng, Yuzhi Wang, Haibin Huang, Donghao Liu, Haoqiang Fan, and Shuaicheng Liu. Nbnet: Noise basis learning for image denoising with subspace projection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4896–4906, 2021.

[9] 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.

[10] Chi-Mao Fan, Tsung-Jung Liu, and Kuan-Hsien Liu. Sunet: swin transformer unet for image denoising. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2333–2337. IEEE, 2022.

[11] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

[12] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2869, 2014.

[13] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020.

[14] 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.

[15] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018.

[16] Xiangyu Li, Yonghong Hou, Pichao Wang, Zhimin Gao, Mingliang Xu, and Wanqing Li. Trear: Transformer-based rgb-d egocentric action recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):246–252, 2021.

[17] 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.

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

[19] Alice Lucas, Michael Iliadis, Rafael Molina, and Aggelos K Katsaggelos. Using deep neural networks for inverse problems in imaging: beyond analytical methods. *IEEE Signal Processing Magazine*, 35(1):20–36, 2018.

[20] 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.

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

[22] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[23] Dongwei Ren, Wei Shang, Pengfei Zhu, Qinghua Hu, Deyu Meng, and Wangmeng Zuo. Single image deraining using bilateral recurrent network. *IEEE Transactions on Image Processing*, 29:6852–6863, 2020.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[25] Uwe Schmidt and Stefan Roth. Shrinkage fields for effective image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2774–2781, 2014.

[26] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[27] 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.

[28] 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.

[29] Chunwei Tian, Lunke Fei, Wenxian Zheng, Yong Xu, Wangmeng Zuo, and Chia-Wen Lin. Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275, 2020.

[30] Chunwei Tian, Yong Xu, Wangmeng Zuo, Bob Zhang, Lunke Fei, and Chia-Wen Lin. Coarse-to-fine cnn for image super-resolution. *IEEE Transactions on Multimedia*, 23:1489–1502, 2020.

[31] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017.

[32] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing*

and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, pages 36–46. Springer, 2021.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[34] 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*, pages 1905–1914, 2021.

[35] 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.

[36] Jun Xu, Lei Zhang, and David Zhang. External prior guided internal prior learning for real-world noisy image denoising. *IEEE Transactions on Image Processing*, 27(6):2996–3010, 2018.

[37] Jun Xu, Lei Zhang, Wangmeng Zuo, David Zhang, and Xiangchu Feng. Patch group based nonlocal self-similarity prior learning for image denoising. In *Proceedings of the IEEE international conference on computer vision*, pages 244–252, 2015.

[38] 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.

[39] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5978–5986, 2019.