Residual Swin Transformer Channel Attention Network for Image Demosaicing

Wenzhu Xing, Karen Egiazarian
Computational Image Group
Tampere University, Tampere, Finland
Email: givenname.surname@tuni.fi

***Abstract*—Image demosaicing is problem of interpolating full-resolution color images from raw sensor (color filter array) data. During last decade, deep neural networks have been widely used in image restoration, and in particular, in demosaicing, attaining significant performance improvement. In recent years, vision transformers have been designed and successfully used in various computer vision applications. One of the recent methods of image restoration based on a Swin Transformer (ST), SwinIR, demonstrates state-of-the-art performance with a smaller number of parameters than neural network-based methods. Inspired by the success of SwinIR, we propose in this paper a novel Swin Transformer-based network for image demosaicing, called RSTCANet. To extract image features, RSTCANet stacks several residual Swin Transformer Channel Attention blocks (RSTCAB), introducing the channel attention for each two successive ST blocks. Extensive experiments demonstrate that RSTCANet outperforms state-of-the-art image demosaicing methods, and has a smaller number of parameters. The source code is available at https://github.com/xingwz/RSTCANet.**

***Index Terms*—Image Demosaicing, Swin Transformer, Channel Attention**

I. Introduction

Most modern digital cameras record only one color channel (red, green, or blue) per pixel. In order to recover the missing pixels, the image demosaicing models are proposed to reconstruct a full color image from a one-channel mosaiced image. Existing demosaicing methods can be classified into two categories: model-based methods [1–4], which recover images based on mathematical models and image priors in the spatial-spectral domain; and learning-based methods [5–11], based on process mapping learned from abundant ground-truth image and mosaiced image pairs. Among these methods, the recent ones [9–11] attain state-of-the-art performance. However, there are still color artifacts in their resulting images (Fig. 1), especially in high frequency regions. Besides, the cost of these networks to improve performance is to increase the depth of the network, which results in a bigger model size (Table IV).

To overcome the above-mentioned problems, we turn our attention to other lighter but effective models for image restoration. Recently proposed vision transformer, called Swin Transformer [12] outperforms state-of-the-art in several vision problems, such as image classification, object detection, and semantic segmentation. Same year, a U-Net method based on Swin Transformer has been proposed for medical image segmentation, called Swin-Unet [13]. Meanwhile, another Swin Transformer-based method, SwinIR [14], was proposed for image restoration. SwinIR surpasses state-of-the-art methods on image super-resolution, image denoising, and JPEG compression artifact reduction with fewer number of parameters. Inspired by the success of SwinIR, we adopt Swin Transformer to propose a lightweight model for image demosaicing. We notice that while utilizing Swin Transformer in SwinIR is helpful to fully excavate the image features patch attentions horizontally, an extraction of the channel features vertically has not received equal attention.

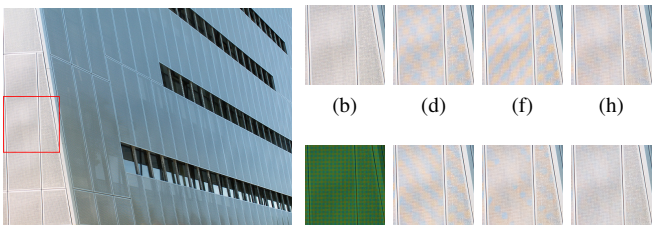


Fig. 1: Visual results comparison of different demosaicing methods on image *026* from Urban100 dataset. (a) Ground-truth and selected area; (b) Ground-truth; (c) Mosaiced; (d) IRCNN; (e) RSTCANet-B; (f) DRUNet; (g) RSTCANet-S; (h) RNAN; (i) RSTCANet-L.

Considering the inter-dependencies among the feature channels should be utilized as well, we introduce the channel attention [15] in the basic block of SwinIR, residual Swin Transformer block (RSTB), to comprehensively extract image features. The proposed combination is named RSTCAB, which is composed of several Swin Transformer layers (STL) and several channel attention blocks. For each two successive STLs, one channel attention block is utilized to generate different attention for each channel-wise feature learned by STLs. The channel attention (CA) is first proposed in RCAN [15]. It consists of a GlobalPooling layer, a down-sampling convolution layer, an up-sampling convolution layer, and the sigmoid function. The ablation study in Sec. IV-A proves the adoption of CA can further improve the performance of RSTB on demosaicing.

There is a recent work on SCUNet [16], a U-Net based on Swin Transformer, for a blind denoising. The basic module SC block of SCUNet combines the Swin Transformer and residual convolutional block. In contrast, our proposed RSTCAB

978-1-6654-6623-3/22/$31.00 © 2022 IEEE

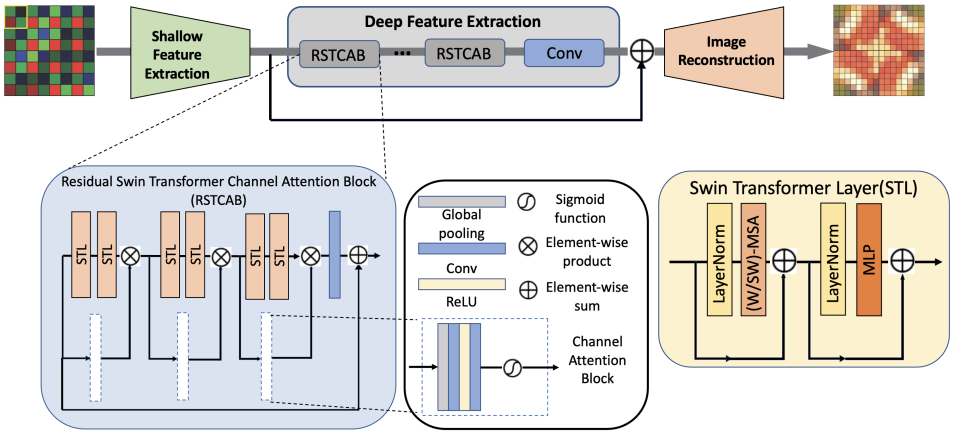


Fig. 2: Residual Swin Transformer Channel Attention Network (RSTCANet) and Residual Swin Transformer Channel Attention Block (RSTCAB).

introduces channel attention blocks in the Swin Transformer blocks.

In summary, there are three main contributions in this paper: (1) We propose the first vision transformer-based method RSTCANet for image demosaicing; (2) The proposed residual Swin Transformer channel attention block (RSTCAB) takes advantage of both Swin Transformer and channel attention. Compared with either other Swin Transformer-based block [14] or residual channel attention block [15], RSTCAB attains the best performance on image demosaicing; (3) RSTCANet achieves state-of-the-art performance on four datasets with smaller model size compared with the existing image demosaicing methods. In addition, the resulting images generated by RSTCANet contain much less visible artifacts (see an example in Fig. 1).

II. Related Works

Vision Transformer. Recently, inspired by the success of Transformer in natural language processing (NLP) domain [17], more and more researchers proposed the Transformer-based architectures for computer vision tasks. The new creation work ViT [18] proposed a Transformer-based architecture for image classification. In order to improve the computational efficiency of Vision Transformer, several works made different efforts, such as the pyramid Transformer architecture [19, 20], and self-attention on local windows [12, 21]. Besides image classification, Transformer-based architectures also achieve impressive performance on other high-level vision tasks, such as object detection [12, 22–24], segmentation [12, 13, 25, 26], and crowd counting [27, 28].

For image restoration tasks, IPT [29] is first proposed based on standard Transformer. However, IPT needs to pretrain on a large-scale synthesized dataset and multi-task learning to get good performance. To improve the efficiency and effectiveness of the Transformer-based on image restoration problems, several methods [14, 30, 31] are proposed in recent years.

III. Method

A. Framework

Network architecture. The architecture of our proposed residual Swin Transformer Channel Attention network (RSTCANet) is shown in Fig. 2. Similar to SwinIR [14], the network consists of three modules: the shallow feature extraction, the deep feature extraction, and the image reconstruction modules. The shallow feature extraction module is composed of a pixel shuffle layer and a vanilla linear embedding layer. For deep feature extraction, we propose residual Swin Transformer Channel Attention blocks (RSTCAB) to extract both hierarchical window based self-attention-aware features [12] and vertical channel-attention-aware features. This module consists of K RSTCAB and one 3×3 convolutional layer. The shallow and deep features are first aggregated by a long skip connection before they fed into the image reconstruction module. The image reconstruction consists of the up-sampling layer and two 3×3 convolutional layers.

Loss function. We optimize the RSTCANet with the $\mathcal{L}_1$ loss function. Given the training pairs $\{I_M^i, I_{GT}^i\}_{i=1}^N$, containing N mosaiced inputs and their corresponding ground truth images, the optimization of the parameters of RSTCANet can be formulated as :

$$\mathcal{L}(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \|RSTCANet(I_M^i) - I_{GT}^i\|_1, \qquad (1)$$

where Θ denotes the parameter set of RSTCANet and $\|\cdot\|_1$ denotes ℓ_1 norm.

B. Residual Swin Transformer Channel Attention Block

As shown in Fig. 2, there are N Swin Transformer layers (STL) and $N/2$ channel attention blocks (CA), and one 3×3 convolutional layer in our proposed residual Swin Transformer Channel Attention block (RSTCAB). There is also a skip connection in the RSTCAB, guaranteeing that the RSTCAB will focus on the differences between the input and output images. The skip connection in RSTCAB and the long skip connection in the network results in the proposed RSTCANet

becoming a residual in residual framework, as in many other image restoration methods [11, 15, 32].

For each two successive STL, the channel attention block generates the channel statistics with the input of two STLs and multiplies the produced attention with the output of two STLs. The N channel attention blocks in the same RSTCAB share parameters. The structure of the channel attention block is same as the one in [15], see Sec. I). With the channel attention blocks, the residual component learned by STL in the RSTCAB is adaptively rescaled. The convolutional layer at the end of RSTCAB is very important for the vision transformer-based image restoration network. We prove this in Sec. IV-A. The discussions about the effect and the number of the channel attention blocks are also included in Sec. IV-A.

C. Architecture Variants

In order to fairly compare our proposed RSTCANet with state-of-the-art image demosaicing methods, we build three different model variants, like in [12]. We build our base model, called RSTCANet-B, to have the smallest model size and computation complexity. We also introduce RSTCANet-S and RSTCANet-L, which are versions of about $3\times$ and $6\times$ the model size and computational complexity, respectively. The parameter settings of these model variants are shown in Table I. The model size of the model variants for image demosaicing is shown in Sec. IV-B.

TABLE I: The parameter settings of different model variants. C is the channel number. K and N denote the number of RSTCAB and the number of STL in one RSTCAB, respectively.

Model Variants	C	K	Multihead	N
RSTCANet-B	72	2	6	6
RSTCANet-S	96	4	6	6
RSTCANet-L	128	4	8	8

IV. Experiments

For the training, we have applied Nvidia Tesla V100 GPU with 32 GB memory from the Tampere University TCSC Narvi computing cluster. We select DIV2K [33] as our training set, which contains 800 training images. Data augmentation is performed on images, which are randomly rotated by $90°$, $180°$, $270°$ and flipped horizontally. The batch size is 16, and the patch size is 64×64. For optimization of the network parameters, we use Adam [34] with $\beta_1 = 0.9, \beta_2 = 0.999$, and the learning rate is initialized to 0.0001. For three model variants, RSTCANet-B,S and L, the learning rate decreases by half each 40k, 100k, and 200k iterations, respectively. Here k equals to 1000. The window size is set to 8 by default. Other parameter settings can be found in Table. I.

A. Ablation Study and Discussion

Table II shows the results of ablation study of RSTCAB. We have investigated the effect of Multihead size (MS), short skip connection (SSC), and the number of channel attention (CA) blocks in one RSTCAB. We have selected the RSTCANet trained with 2 RSTCAB, Multihead size 4, three channel attention blocks in one RSTCAB, channel number (C) 64 as the benchmark (see Table II). All models are evaluated for image demosaicing on McM dataset [35] by two metrics, cPSNR and SSIM.

TABLE II: The ablation study of different components of RSTCAB.

Case	MS	SSC	CA	cPSNR/SSIM
RSTCANet	4	✗	3	38.71/0.9897
RSTCANet-h2	2	✗	3	38.55/0.9892
RSTCANet-SSC	4	✓	3	38.60/0.9896
RSTCANet-CA0	4	✗	0	38.66/0.9897
RSTCANet-CA1	4	✗	1	38.68/0.9896
RSTCANet-CA6	4	✗	6	38.65/0.9896

Impact of Multihead size. We have designed a RSTCANet-h2, with size of Multihead equals to 2. It can be observed that the model performance can be improved by a bigger attention Multihead size with the same channel number.

Impact of short skip connection. The RSTCANet-SSC is designed to check if adding a short skip connection for every two successive STLs would provide any improvement. By comparing RSTCANet and RSTCANet-SSC, one can see that extra skip connection reduces the performance of RSTCAB.

The impact of the number of channel attention blocks in one RSTCAB. Three other variations of RSTCANet are designed to examine the effect of CA blocks. There are no CA blocks in RSTCANet-CA0. Note that the structure of RSTCANet-CA0 is identical to the structure of SwinIR [14]. In the RSTCAB of RSTCANet-CA1 there is only one CA block, and the input of this CA block is the input of RSTCAB. The attention generated by this CA block is multiplied by the features produced by the sixth STL of RSTCAB. For RSTCANet-CA6, in the RSTCAB, there is one CA block for each STL.

A comparison with RSTCANet-CA0 presented in Table II, shows that by exploiting one CA block for six STLs (RSTCANet-CA1) or applying one CA block for each two successive STLs (RSTCANet) can improve the performance of the RSTCAB. It also demonstrated that the proposed RSTCANet outperforms another Swin Transformer-based method SwinIR on image demosaicing. However, when there are six CA blocks in RSTCAB (RSTCANet-CA6), a performance of RSTCANet becomes worse, which can be explained by the shifted window partitioning mechanism for two successive STLs [12]. To make up for the deficiency of cross-window connections in the window-based self-attention module, the authors of [12] introduced the shifted window partitioning strategy in two successive Swin transformer blocks. When the channel attention is learned for each STL, the connections across windows are ignored. In contrast, by applying the channel attention for every two successive STLs, there is a positive effect of the shifted window partitioning strategy.

Impact of Convolutional layers. We have also added extra convolution layers in the RSTCANet, at the end of

TABLE III: The ablation study of convolutional layers in RSTCANet. Image demosaicing on McM dataset. #Conv. represent the number of convolution layers. DFE denotes the deep feature extraction module.

Case	#Conv. in RSTCAB	#Conv. in DFE	cPSNR/SSIM	#param. (MB)
RSTCANet-B	1	1	38.89/0.9902	5.5
RSTCANet-1	1	2	38.88/0.9899	5.7
RSTCANet-2	2	1	38.82/0.9898	5.9
RSTCANet-3	0	1	38.52/0.9894	5.1

RSTCAB and at the end of deep feature extraction module. We have trained these two variations of RSTCANet, and denoted them as RSTCANet-1 and RSTCANet-2, respectively. Another variation, RSTCANet-3 without convolutions in the RSTCAB is trained as well. All models in this experiment are trained with channel features 72 and Multihead size 6 as RSTCANet-B. From the results presented in Table III, we unexpectedly find that utilization of more convolution layers in RSTCANet does not guarantee the performance improvement, but leads to an increase of the model size.However, the experimental result demonstrated (RSTCANet-3) that the convolutional layer at the end of RSTCAB is necessary. It should be the case because the STL layers are more relevant for recognition rather than reconstruction [39].

Impact of the Basic Block. A comparison between RSTCANet and RSTCANet-CA0 in Table II shows that RSTCAB performs better than RSTB [14] on demosaicing. Besides this, we train another RSTB-based model with a bigger channel number (72), denoted as SwinIR*. We also compare our proposed RSTCAB with other two related basic blocks, RCAB [15] and SC block [16]; we marked these models as RCAN* and SCNet*, respectively. The number of RCAB in RCAN* is 24, and the number of SC blocks in SCNet* is 2 and each SC block has 6 Swin Transformer blocks. For fair comparison, all other training settings of SwinIR*, RCAN* and SCNet* are same with the proposed RSTCANet-B, including the structure of shallow feature extraction module and the image reconstruction module.

From Table V, one can see that combining the CA blocks whose parameters are shared with the Swin Transformer blocks can improve the demosaicing performance of model without additional storage cost. The numbers of parameters of RSTCANet-B and SwinIR* are same, but the demosaicing performance is improved by 0.13 dB. In addition, the model size of RCAN* is 1.4 times of RSTCANet-B. While the number of parameters is more, the demosaicing performance of RCAN* on McM set is slightly worse than RSTCANet-B. This shows that the proposed RSTCAB combines advantages of RCAB and RSTB. For SCNet*, it achieves a slight improvement (0.11 dB) at the cost of 1.4 times parameters of RSTCANet-B. This demonstrates that RSTCAB has a better trade-off between the demosaicing performance and model size than SC block. We also report FLOPs and run-time comparisons in Table V. Among these basic blocks, RSTCAB achieves the best trade-off between performance, FLOPs, runtime and #param.

B. Results on Image Demosaicing

Quantitative comparison. Table IV shows the quantitative comparisons between RSTCANet and the state-of-the-art methods: DRUNet [9], IRCNN [10] and RNAN [11]. We test different methods on four benchmark datasets, McM [35], Kodak [36], CBSD68 [37] and Urban100 [38]. The color PSNR and SSIM values are evaluated on resulting images.

As one can see, compared with these state-of-the-art demosaicing methods, our RSTCANet can get comparable performance with smaller size. The RSTCANet-B performs better than IRCNN at least 1 dB with 0.3× model size. RSTCANet-S outperforms DRUNet by 0.08 dB on McM, 0.31 dB on Kodak and 0.47 dB on Urban100 with only 0.13× of its size. On CBSD68, our RSTCANet-S also performs slightly better than DRUNet. Compared with the SOTA method RNAN, our RSTCANet-S gets the comparable performance with only half size parameters. On Urban100, the RSTCANet-S even achieves a slightly better performance (0.04 dB) than RNAN.

The large model variant, RSTCANet-L, is a lighter (1.7 MB) than RNAN, but has better performance than RNAN on three datasets. Especially on Urban100, our RSTCANet-L performs better than RNAN by 0.42 dB.

In addition, by increasing the channel number C and the number of RSTCAB blocks, RSTCANet-S and RSTCANet-L improve the performance by at least 0.5 dB and 0.63 dB compared with RSTCANet-B.

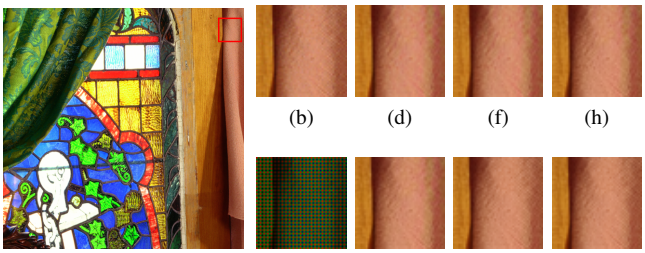


Fig. 3: Visual comparison of different demosaicing methods on image *1* from McM dataset. (a) Full ground-truth and selected area; (b) Ground-truth; (c) Mosaiced; (d) IRCNN; (e) RSTCANet-B; (f) DRUNet; (g) RSTCANet-S; (h) RNAN; (i) RSTCANet-L.

Visual comparison. Fig. 1 and Fig. 3-4 illustrate the visual comparisons between our proposed RSTCANet and the state-of-the-art demosaicing methods. In these illustrations, one can observe that the proposed RSTCANet-L can generate less color artifacts than other methods. For high frequency regions, the color artifacts exist even in RNAN resulting images, such as the right border of the curtain in Fig. 3, and the texture of jeans in Fig. 4. In contrast, our method can reconstruct the color image with less color artifacts.

V. Conclusion

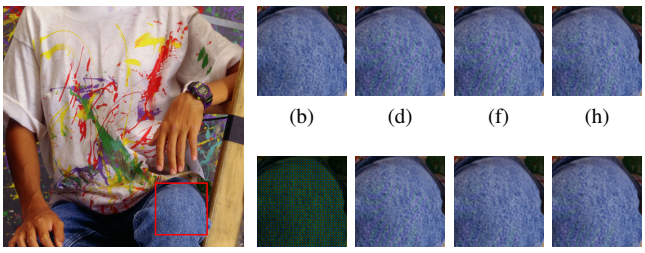
In this paper, we propose a Swin Transformer-based image demosaicing model RSTCANet, based on the residual Swin

TABLE IV: The quantitatively comparison with state-of-the-art methods for demosaicing on benchmark datasets. The last column is the size of the model. The best values are in **bold**.

Method	McM [35] cPSNR/SSIM	Kodak [36] cPSNR/SSIM	CBSD68 [37] cPSNR/SSIM	Urban100 [38] cPSNR/SSIM	Size (MB)
IRCNN [10]	37.84/0.9885	40.65/0.9915	40.31/0.9924	37.03/0.9864	18.0
DRUNet [9]	39.40/0.9914	42.30/0.9944	42.33/0.9955	39.22/0.9906	124.5
RNAN [11]	39.66/0.9915	**42.92/0.9952**	42.45/0.9959	39.65/0.9923	34.3
RSTCANet-B	38.89/0.9902	42.11/0.9948	41.74/0.9954	38.52/0.9906	5.5
RSTCANet-S	39.58/0.9910	42.61/0.9951	42.36/0.9958	39.69/0.9924	16.0
RSTCANet-L	**39.91/0.9916**	42.74/**0.9952**	**42.47/0.9960**	**40.07/0.9931**	32.6

TABLE V: The ablation study of basic block of demosaicing network. Image demosaicing on McM dataset (500×500).

Case	Basic Block	cPSNR/SSIM	#param.(MB)	Runtime	FLOPs
RSTCANet-B	RSTCAB	38.89/0.9902	5.5	0.375s	93.5G
SwinIR*	RSTB [14]	38.76/0.9898	5.5	0.368s	92.9G
RCAN*	RCAB [15]	38.86/0.9899	7.7	0.308s	52.8G
SCNet*	SC block [16]	39.00/0.9901	8.5	0.385s	200.8G



(a) McM:5.png (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 4: Visual comparison of different demosaicing methods on image *5* from McM dataset. (a) Full ground-truth and selected area; (b) Ground-truth; (c) Mosaiced; (d) IRCNN; (e) RSTCANet-B; (f) DRUNet; (g) RSTCANet-S; (h) RNAN; (i) RSTCANet-L.

Transformer Channel Attention blocks (RSTCAB), which takes advantage of both Swin Transformer and Channel Attention blocks. Experimental results show that RSTCAB surpass other Swin Transformer-based blocks on image demosaicing. The quantitative and qualitative results also demonstrate that RSTCANet achieves state-of-the-art performance on image demosaicing, generating much less color artifacts in the resulting images. In the future, we plan to extend the RSTCANet to other image restoration tasks, such as image denoising and super-resolution.

REFERENCES

[1] K. Hirakawa and T. W. Parks, "Adaptive homogeneity-directed demosaicing algorithm," *IEEE Transactions on Image Processing*, vol. 14, no. 3, pp. 360–369, 2005.

[2] H. S. Malvar, L.-w. He, and R. Cutler, "High-quality linear interpolation for demosaicing of bayer-patterned color images," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3. IEEE, 2004, pp. iii–485.

[3] C.-Y. Su, "Highly effective iterative demosaicing using weighted-edge and color-difference interpolations," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 2, pp. 639–645, 2006.

[4] L. Zhang and X. Wu, "Color demosaicking via directional linear minimum mean square-error estimation," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2167–2178, 2005.

[5] F.-L. He, Y.-C. F. Wang, and K.-L. Hua, "Self-learning approach to color demosaicking via support vector regression," in *2012 19th IEEE International Conference on Image Processing*. IEEE, 2012, pp. 2765–2768.

[6] J. Sun and M. F. Tappen, "Separable markov random field model and its applications in low level vision," *IEEE transactions on image processing*, vol. 22, no. 1, pp. 402–407, 2012.

[7] N.-S. Syu, Y.-S. Chen, and Y.-Y. Chuang, "Learning deep convolutional networks for demosaicing," *arXiv preprint arXiv:1802.03769*, 2018.

[8] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand, "Deep joint demosaicking and denoising," *ACM Transactions on Graphics (ToG)*, vol. 35, no. 6, pp. 1–12, 2016.

[9] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, "Plug-and-play image restoration with deep denoiser prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[10] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep cnn denoiser prior for image restoration," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3929–3938.

[11] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," *arXiv preprint arXiv:1903.10082*, 2019.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[13] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint*

arXiv:2105.05537, 2021.

[14] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.

[15] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286–301.

[16] K. Zhang, Y. Li, J. Liang, J. Cao, Y. Zhang, H. Tang, R. Timofte, and L. Van Gool, "Practical blind denoising via swin-conv-unet and data synthesis," *arXiv preprint arXiv:2203.13278*, 2022.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 936–11 945.

[20] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[21] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 894–12 904.

[22] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[23] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, no. 2, pp. 261–318, 2020.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[25] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," *arXiv preprint arXiv:2006.03677*, 2020.

[26] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[27] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai, "Transcrowd: weakly-supervised crowd counting with transformers," *Science China Information Sciences*, vol. 65, no. 6, pp. 1–14, 2022.

[28] G. Sun, Y. Liu, T. Probst, D. P. Paudel, N. Popovic, and L. Van Gool, "Boosting crowd counting with transformers," *arXiv preprint arXiv:2105.10926*, 2021.

[29] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.

[30] Z. Wang, X. Cun, J. Bao, and J. Liu, "Uformer: A general u-shaped transformer for image restoration," *arXiv preprint arXiv:2106.03106*, 2021.

[31] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," *arXiv preprint arXiv:2111.09881*, 2021.

[32] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2480–2495, 2020.

[33] E. Agustsson and R. 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.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] L. Zhang, X. Wu, A. Buades, and X. Li, "Color demosaicking by local directional interpolation and nonlocal adaptive thresholding," *Journal of Electronic imaging*, vol. 20, no. 2, p. 023016, 2011.

[36] E. Kodak, "Kodak lossless true color image suite (photocd pcd0992)," *URL http://r0k. us/graphics/kodak*, vol. 6, 1993.

[37] D. Martin, C. Fowlkes, D. Tal, and J. 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*, vol. 2. IEEE, 2001, pp. 416–423.

[38] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5197–5206.

[39] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *arXiv preprint arXiv:2111.06377*, 2021.