Published in Image Processing On Line on 2022–12–20.
Submitted on 2022–10–10, accepted on 2022–10–10.
ISSN 2105–1232 © 2022 IPOL & the authors CC–BY–NC–SA
This article is available online with supplementary materials,
software, datasets and online demo at
https://doi.org/10.5201/ipol.2022.430

A Brief Analysis of the SwinIR Image Super-Resolution

Ngoc-Long Nguyen

Université Paris-Saclay, ENS Paris-Saclay, Centre Borelli, Gif-sur-Yvette, France
ngoc_long.nguyen@ens-paris-saclay.fr

Communicated by Jean-Michel Morel *Demo edited by* Ngoc Long Nguyen

Abstract

SwinIR is a recent image restoration method based on the Swin Transformer architecture. In contrast to other traditional convolutional neural networks, SwinIR is capable of capturing sophisticated attention between image patches, leading to remarkable results. In this paper, we focus on the aspect of single-image super-resolution by SwinIR. We discuss the characteristics of the architecture of this algorithm and compare it to other deep learning methods.

Source Code

The source code and documentation for this algorithm are available from the web page of this article[1]. Usage instructions are included in the `README` file of the archive. The original implementation of the method is available here[2].
This is an MLBriefs article, the source code has not been reviewed!

Keywords: single image super-resolution; vision transformer; Swin transformer

1 Introduction

Single image super-resolution (SISR) is a fundamental problem in computer vision that aims to obtain a high resolution (HR) output from its degraded low-resolution (LR) counterpart. Recently, deep-learning methods have outperformed traditional SISR algorithms by a huge margin in both quantitative and qualitative results. As a matter of fact, SISR can be seen as an interpolation problem since SISR tries to recover pixels in the HR from their neighboring pixels in the LR image. Being a local problem, SISR has been dominated by convolutional neural networks (CNN). In particular, [13] uses dilated convolution to increase the receptive field over twice and get better results. RDN [15] proposes a residual dense network to exploit the hierarchical features from the convolutional layers. RCAN [14] adds an attention mechanism inside the CNN framework to exploit better feature representation produced by the channels.

[1] https://doi.org/10.5201/ipol.2022.430
[2] https://github.com/JingyunLiang/SwinIR

On the other side of deep-learning, Transformer is the backbone of natural language processing (NLP). Since its invention in 2017, Transformer with its powerful self-attention mechanism has refreshed and dominated all modern architectures in NLP. The question we all wanted to ask is whether Transformer could be applicable to computer vision. One naive approach is to consider image pixels as tokens and put all of them into the self-attention mechanism. However, this approach is intractable due to the enormous amount of pixels in natural images. To this aim, Dosovitskiy et al. [3] introduce the Vision Transformer (ViT) which applies Transformer directly on non-overlapping image patches. Achieving state-of-the-art performance in image classification, ViT is very promising in computer vision. Notwithstanding its great potential, the limitation of the ViT resides in its quadratic computational complexity on image size, which makes it unscalable to higher-resolution images. Another related work IPT [2] uses a pretrained Transformer model to perform image processing tasks. Like other Vision Transformer-based models, IPT is computationally intensive and requires a large training dataset. Liu et al. [6] propose the Swin Transformer to overcome the main drawbacks of the Vision Transformer and achieve state-of-the-art results in image classification, object detection, and semantic segmentation. The Swin Transformer alleviates the computational burden of the ViT by computing self-attention only locally, but also models long-range dependency by using the shifted window scheme. The Swin Transformer is used in many state-of-the-art super-resolution methods, including stereo image super-resolution [4] and burst raw super-resolution [9].

Recently, Liang et al. [5] proposed SwinIR, an excellent baseline for image restoration based on the Swin Transformer. SwinIR is actually a hybrid model with two CNN modules (shallow feature extraction and high-quality image reconstruction) at the two ends, and specially a Swin Transformer-based module (deep feature extraction) as the crucial component of the method. SwinIR is proven to achieve state-of-the-art performance on single image super-resolution, image denoising, and JPEG artifact removal with a reasonable number of parameters. In this project, we analyze the performance of SwinIR on SISR and examine whether long-range information by the Transformer is beneficial to such a local problem.

2 Method

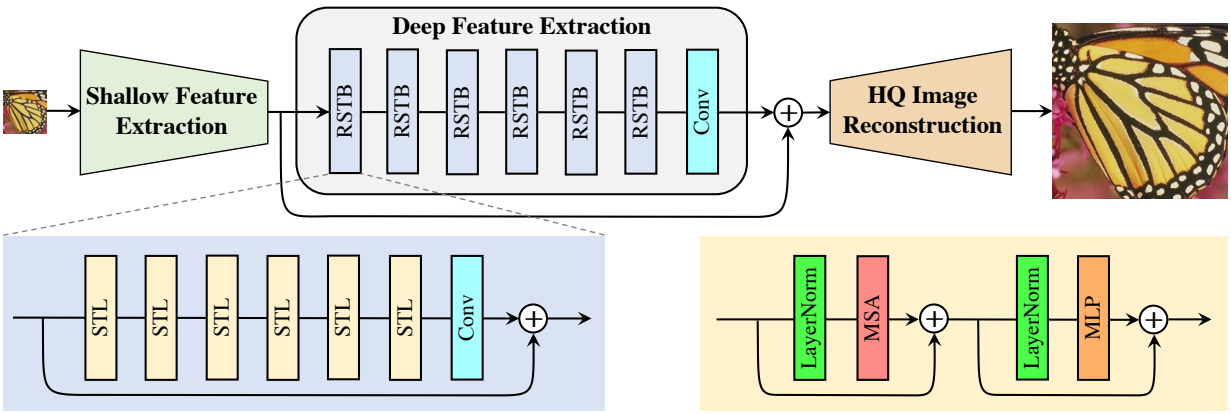
2.1 SwinIR Architecture

As shown in Figure 1, SwinIR has a hybrid architecture consisting of three modules: shallow feature extraction (CNN), deep feature extraction (Swin Transformer), and high quality image reconstruction (CNN). In this project, we focus on two SwinIR networks: *classical SR* and *realistic SR*. The classical model is a medium-sized network designed and trained for quantitative measurement. The realistic model is larger and is trained to perform real-world super-resolution.

Shallow feature extraction. The shallow feature extraction can be considered a preprocessing step, which serves to map the LR image $I_{LR} \in \mathbb{R}^{H\times W\times 3}$ to a richer dimensional feature space with C feature channels. The shallow feature extraction is a convolutional layer H_{SF} with kernel size 3×3

$$F_0 = H_{SF}(I_{LR}), \tag{1}$$

where $F_0 \in \mathbb{R}^{H\times W\times C}$ is the shallow extracted feature. Applying an early small convolutional layer at the beginning of the Vision Transformer was reported to help the training to stabilize and converge faster [11]. The embedded dimension C is set to 180 for the classical model and 240 for the realistic model.



(a) Residual Swin Transformer Block (RSTB) (b) Swin Transformer Layer (STL)

Figure 1: SwinIR architecture.

Deep feature extraction. The deep feature extraction, composed of K residual Swin Transformer blocks (RSTB) and a CNN, comes after the shallow layer H_{SF}. The value of K is set to 6 in the classical model and 9 in the realistic model. Concretely, first these blocks RSTB compute the transitional features $F_1, F_2, \ldots, F_K$ sequentially

$$F_i = H_{RSTB_i}(F_{i-1}), i = 1, 2, \ldots, K, \tag{2}$$

where H_{RSTB_i} denotes the $i-$th RSTB. And then a small CNN H_{CONV} at the end extracts the output deep feature F_{DF}

$$F_{DF} = H_{CONV}(F_K). \tag{3}$$

This CNN is presumed to introduce the image domain-specific inductive biases into the Transformer. The CNN in the classical model is just a simple convolutional layer that keeps the embedded dimension $C = 180$. For the realistic model, it is an hourglass-shaped CNN with 3 convolutional layers and hidden dimension 60 in order to save parameters and memory.

High resolution image reconstruction. Finally, the reconstruction module H_{REC} produces the high resolution output from the computed shallow and deep features,

$$I_{HR} = H_{REC}(F_0 + F_{DF}). \tag{4}$$

The shallow features and the deep features contain mainly the low-frequency and the high-frequency information, respectively. While the former is pretty simple to extract with a convolutional layer, the latter is much more sophisticated to reconstruct. Hence, a long skip connection from F_0 up to F_{DF} is used to help the deep feature extraction focus on recovering the high frequency details. The reconstruction module H_{REC} is built from an upsample operator (pixel shuffle [10] in the classical model, and nearest-neighbor interpolation in the realistic model) and several convolution layers.

2.2 Residual Swin Transformer Block

Each residual Swin Transformer block (RSTB) is composed of L ($L = 6$ for the two models) Swin Transformer Layers (STL) followed by a CNN (Figure 1a). More specifically, given the input feature $F_{i,0}$ of the $i-$th RSTB, the intermediate features $F_{i,1}, \ldots, F_{i,L}$ by L STL are computed as

$$F_{i,j} = H_{STL_{i,j}}(F_{i,j-1}), j = 1, 2, \ldots, L, \tag{5}$$

where $H_{STL_{i,j}}$ is the $j-$th STL of the $i-$th RSTB. Then a CNN is applied to enhance the translation equivariance of the Swin Transformer just before the residual connection

$$F_{i,out} = H_{CONV_i}(F_{i,L}) + F_{i,0}, \tag{6}$$

where H_{CONV_i} is the CNN in the $i-$th RSTB. Note that this CNN has the same architecture as the H_{CONV} in (3). The residual connection stabilizes the training and allows the accumulation of the features at different depths. It is worth noticing that unlike the original Swin Transformer architecture, in the RSTB there is no patch-merging operation (i.e., combine 2×2 image patches into a larger patch) between STL. Moreover, the embedded dimension is kept constant through the layers.

Swin Transformer layer. The Swin Transformer Layer (Figure 1b) has the same structure as in [6]. Basically, first the input features of a STL are partitioned into non-overlapping $M \times M$ local windows ($M = 8$ pixels). Then the standard multi-head self-attention (MSA) is computed for the patches in each window. The number of heads h is fixed to 6 in the classical model and 8 in the realistic model. Next a multi-layer perceptron (MLP) with 2 connected layers (the hidden dimension is two times the embedded dimension C) and GELU non-linearity is used for further feature transformation. LayerNorm (LN) is applied before both MSA and MLP, and the residual connection is applied after both modules. The whole process is then repeated but with the shifted window mechanism (that is, by cyclic shifting the windows by $\frac{M}{2}$ in each direction) to enable cross-window connections.

3 Training Details

3.1 Training Set

For the classical model, the authors use two datasets DIV2K[3] (800 images), and DIV2K + Flickr2K[4] (2650 images), with bicubic downsampling to create training sets. They observe that the model trained with more data has better PSNR performance (+0.3dB) when tested on the dataset Manga109[5]. On the other hand, a large collection of diverse datasets (DIV2K +Flickr2K + OST[6] (10324 images, nature) + WED[7](4744 images) + FFHQ[8] (first 2000 images, face) + Manga109 (manga) + SCUT-CTW1500[9] (first 100 images, texts)) are used to train the realistic model. Furthermore, a sophisticated degradation model from [12] is adopted to simulate real-world scenarios.

3.2 Training Loss and Optimization

The classical model is trained with a simple L_1 loss, while the realistic model is trained with a combination of L_1 loss, GAN loss, and perceptual loss to obtain better visual quality. The two models are both trained 10^6 epochs on 8 GPUs. They are optimized using Adam solver (initial learning rate $= 1e-4$) and MultiStepLR learning rate scheduler with 5 steps and $\gamma = 0.5$. The batch size is set to 32 and the LR image size is 64×64 pixels.

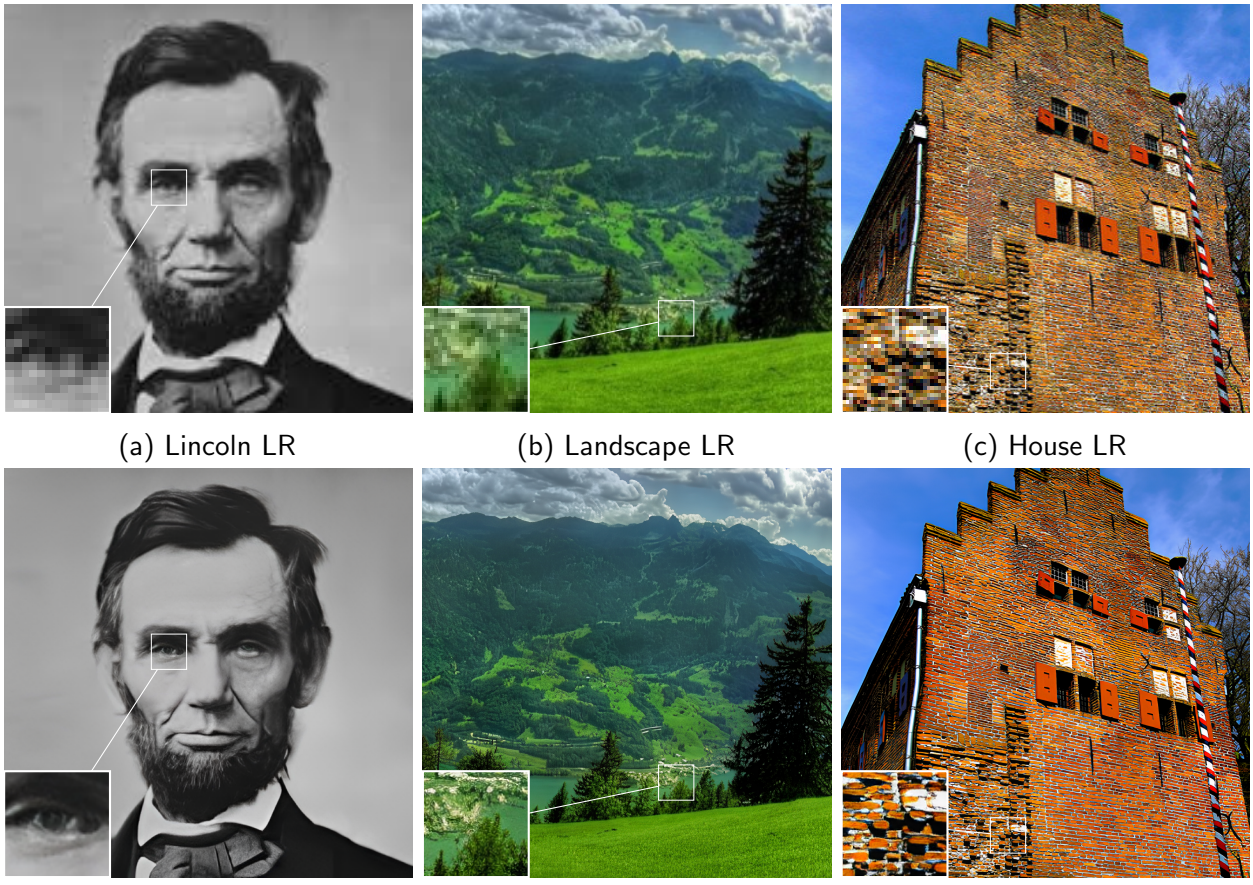
[3] https://cv.snu.ac.kr/research/EDSR/DIV2K.tar
[4] https://cv.snu.ac.kr/research/EDSR/Flickr2K.tar
[5] https://drive.google.com/file/d/13NsteslsUnPj6_Z4wKJg9J1i3eXokCyC/view
[6] https://drive.google.com/drive/folders/1iZfzAxAwOpeutz27HC56_y5RNqnsPPKr
[7] http://ivc.uwaterloo.ca/database/WaterlooExploration/exploration_database_and_code.rar
[8] https://drive.google.com/drive/folders/1tZUcXDBeOibC6jcMCtgRRz67pzrAHeHL
[9] https://universityofadelaide.box.com/shared/static/py5uwlfyyytbb2pxzq9czvu6fuqbjdh8.zip

4 Experiments

In the demo, we fix the super-resolution factor to 4 since the authors only provide the x4 pretrained model for the realistic SwinIR.

4.1 Real-World Super-Resolution

This section presents the qualitative performance of the realistic model on real-world images. Note that the realistic model is trained to perform not only super-resolution but also image denoising and JPEG artifacts removal, which makes it particularly suitable to restore old pictures or to enhance the quality of natural images. Figure 2 shows the super-resolution reconstruction of the realistic model on real images. Generally, SwinIR excels at removing noise, JPEG artifacts, and producing plausible high-frequency details. But we also notice that when dealing with highly compressed or very noisy images, SwinIR may present unwanted artifacts such as residual noise or cartooned textures, respectively.



(a) Lincoln LR (b) Landscape LR (c) House LR

(d) Lincoln SR (e) Landscape SR (f) House SR

Figure 2: Visual quality of SwinIR super-resolution on real-world images. Top line corresponds to the LR input. Bottom line shows the x4 SR reconstruction of SwinIR.

4.2 Auto-Similarity and Single Image Super-Resolution

We know that SISR is a local problem per se. In this experiment, we want to study the impact of self-attention in the Swin Transformer in the Urban100 dataset. We choose this dataset because it contains a lot of auto-similar structures. The importance of auto-similarity in image restoration was

first exploited in the Non-Local Means denoising method [1]. The authors of [1] demonstrate that we can reduce the noise of an image patch by aggregating its similar patches, which are not necessarily spatially close to the patch of interest. Since this work, auto-similarity has become more and more popular in image processing. Recently, ESRT [8] exploits auto-similarity to train a Transformer network for single-image super-resolution. It is arguable that Transformer (and Swin Transformer) can make use of attention in similar patches to get a better SR reconstruction, especially in the low-contrast or aliased regions. We also compare these Transformer networks with RCAN [14], a classic CNN for SISR.

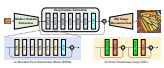
We use bicubic interpolation to create low-resolution images from the Urban100 dataset. Note that we do not include recent GAN-based state-of-the-art methods in this study since they will hallucinate low-contrast details. Both ESRT and RCAN are trained with L_1 loss on the DIV2K dataset with bicubic degradation. The classic Swin Transformer model is trained with L_1 loss but on the DIV2K + Flickr2K dataset.

Figure 3 shows the comparison between the two SwinIR models, RCAN and ESRT on the Urban100 dataset. First, we observe that the SwinIR realistic model is not reliable for recovering the true details due to its generative nature (Figure 3d). Second, we expected ESRT to perform better on this particular test set using global attention (compare for example, Figure 3b and Figure 3c). Maybe the performance of ESRT is restricted by its capacity (ESRT is a lightweight network). Finally, the classical SwinIR network recovers genuinely the low-contrast and aliased textures and achieves the best results. Unfortunately, we could not claim whether this boost of performance comes from the long-range dependency mechanism. First, the window size of SwinIR is really small (8 pixels), which makes SwinIR a rather local network. Second, the classical SwinIR is trained on a larger dataset. Finally, this gain in performance may be due to the advance in network design (i.e., large kernel size, GELU activation, Layer norm, etc) rather than the superiority of Transformer over traditional CNN [7]. In conclusion, SwinIR is a promising and powerful method for SISR, but we still need to carry out more experiments to fully understand its competence.

5 Conclusion

In this paper, we analyzed SwinIR – a Swin Transformer network for image super-resolution. We also discussed how Transformer would benefit from auto-similarity in natural images to get a better performance in SISR. Overall, SwinIR achieves remarkable performance in both synthetic and real-world images, and it has great potential to become a popular backbone in computer vision.

Image Credits

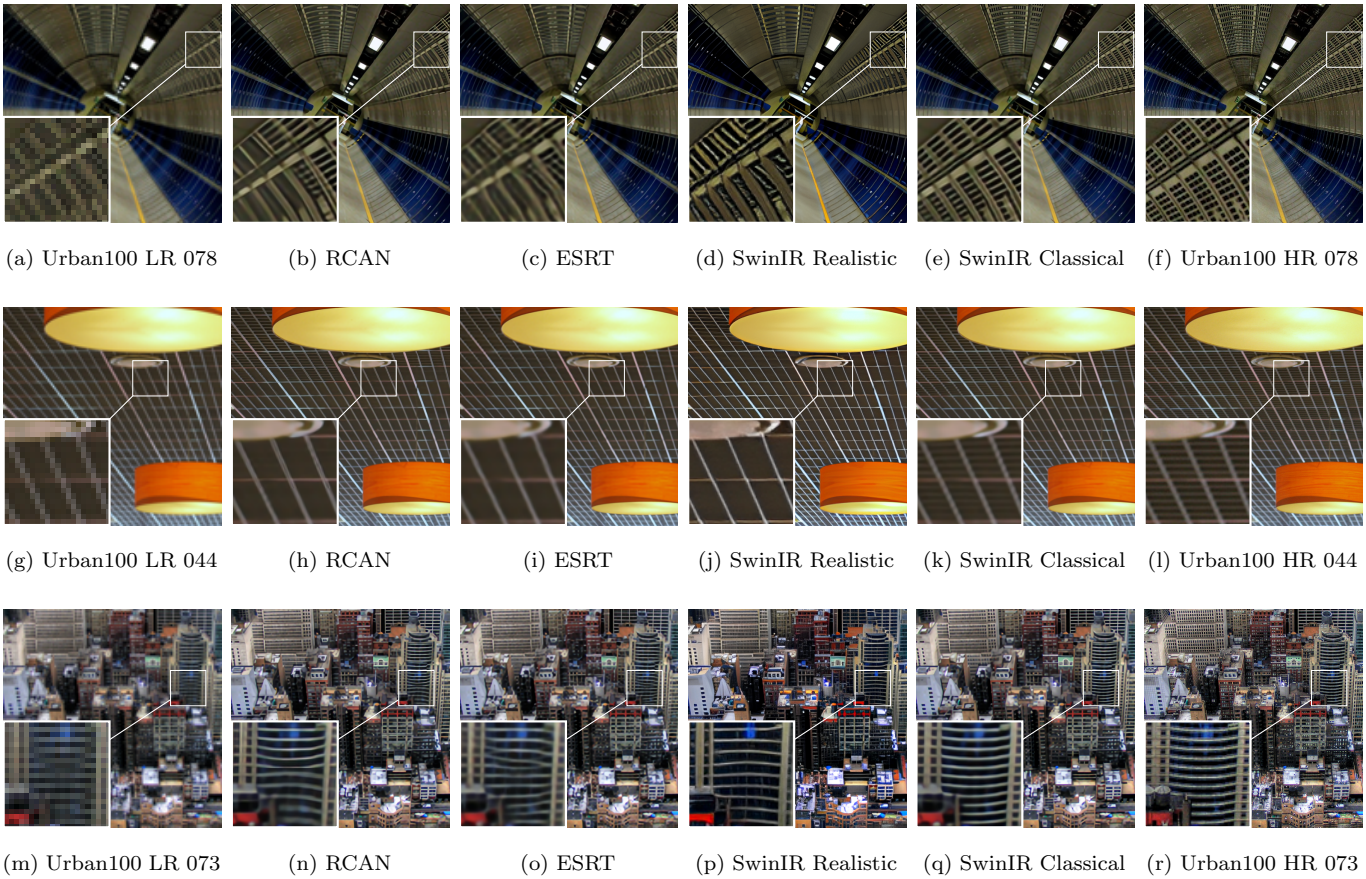

SwinIR[10] original architecture


Dataset RealSRSet+5images[11]


Internet Media[12]


Dataset Urban100[13]

[10] https://arxiv.org/abs/2108.10257
[11] https://github.com/JingyunLiang/SwinIR/releases/download/v0.0/RealSRSet+5images.zip
[12] https://www.zastavki.com/eng/Nature/Nature/Seasons/Summer/wallpaper-78337-12.htm
[13] https://drive.google.com/file/d/1UlNulSoyflrEObwu19BBlT7f2_Wxycga/view?usp=sharing



(a) Urban100 LR 078 (b) RCAN (c) ESRT (d) SwinIR Realistic (e) SwinIR Classical (f) Urban100 HR 078

(g) Urban100 LR 044 (h) RCAN (i) ESRT (j) SwinIR Realistic (k) SwinIR Classical (l) Urban100 HR 044

(m) Urban100 LR 073 (n) RCAN (o) ESRT (p) SwinIR Realistic (q) SwinIR Classical (r) Urban100 HR 073

Figure 3: Qualitative comparison between the two SwinIR models, RCAN, and ESRT on the Urban100 dataset. Super-resolution by factor of 4.

References

[1] A. Buades, B. Coll, and J-M. Morel, *Non-local means denoising*, Image Processing On Line, 1 (2011), pp. 208–212. https://doi.org/10.5201/ipol.2011.bcm_nlm.

[2] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12299–12310. http://dx.doi.org/10.1109/CVPR46437.2021.01212.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2020. https://doi.org/10.48550/arXiv.2010.11929.

[4] K. Jin, Z. Wei, A. Yang, S. Guo, M. Gao, X. Zhou, and G. Guo, *SwiniPASSR: Swin Transformer Based Parallax Attention Network for Stereo Image Super-Resolution*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022, pp. 920–929. https://doi.org/10.1109/CVPRW56347.2022.00106.

[5] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, *SwinIR: Image Restoration using Swin Transformer*, in IEEE/CVF International Conference on Computer

Vision Workshops (ICCVW), 2021, pp. 1833–1844. https://doi.ieeecomputersociety.org/10.1109/ICCVW54120.2021.00210.

[6] 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*, in IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022. https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00986.

[7] Z. Liu, H. Mao, C-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, *A ConvNet for the 2020s*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11976–11986. https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01167.

[8] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, *Transformer for single image super-resolution*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022, pp. 457–466. https://doi.ieeecomputersociety.org/10.1109/CVPRW56347.2022.00061.

[9] Z. Luo, Y. Li, S. Cheng, L. Yu, Q. Wu, Z. Wen, H. Fan, J. Sun, and S. Liu, *BSRT: Improving Burst Super-Resolution with Swin Transformer and Flow-Guided Deformable Alignment*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022, pp. 998–1008. https://doi.ieeecomputersociety.org/10.1109/CVPRW56347.2022.00113.

[10] W. Shi, J. Caballero, F. Huszár, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, *Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1874–1883. https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.207.

[11] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, *Early convolutions help transformers see better*, Advances in Neural Information Processing Systems, 34 (2021), pp. 30392–30400. https://doi.org/10.48550/arXiv.2106.14881.

[12] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, *Designing a practical degradation model for deep blind image super-resolution*, in IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4791–4800. https://doi.org/10.48550/arXiv.2103.14006.

[13] K. Zhang, W. Zuo, S. Gu, and L. Zhang, *Learning deep CNN denoiser prior for image restoration*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3929–3938. https://doi.org/10.1109/CVPR.2017.300.

[14] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, *Image super-resolution using very deep residual channel attention networks*, in European Conference on Computer Vision (ECCV), 2018, pp. 286–301. https://doi.org/10.1007/978-3-030-01234-2_18.

[15] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, *Residual dense network for image super-resolution*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2472–2481. http://dx.doi.org/10.1109/CVPR.2018.00262.