Arbitrary Scale Upsamling with Hybrid Attention and Fourier Analysis

https://github.com/cankinik/Super-Resolution-LTE-HAT

Can KINIK
cankinik@umich.edu
University of Michigan, Ann Arbor

Chia-Ching Tsai
cctsaitw@umich.edu
University of Michigan, Ann Arbor

Abstract

Transformers have shown remarkable performance in the classical computer vision problem of single image super-resolution (SISR). However, recent models come with shortcomings in the areas of having limited capability in spatial range utilization, and requiring separate models to be trained and stored for different upscaling values. These problems have been separately addressed in two recent papers by utilizing a hybrid attention model to activate more input pixels, and using multi-layer perceptrons (MLP) to leverage the underlying implicit function when upsampling to produce arbitrary resolution respectively. Since both papers utilize the same transformer backbone, we propose a novel network that incorporates the strengths of both architectures. Comparisons between the standalone baseline and the proposed hybrid model demonstrate that our methodology increases the PSNR by roughly ***0.06dB****.*

1. Introduction

SISR is one of the most fundamental tasks in low-level computer vision, where a lower-resolution (LR) input image is used to reconstruct a higher-resolution (HR) counterpart [2, 5, 14, 22, 43, 48, 49, 50, 51]. It has many applications, ranging from improving the quality of existing images that were taken under constrained conditions such as CCTV surveillance footage [60], to increasing the frames per second performance of graphical computation by rendering models in lower resolution and then upsampling [46].

Up until recently, the dominant approach has been to use convolutional deep neural networks to extract feature maps, and then upsampling these features through the use of sub-pixel convolution to arrive at a higher resolution output [25, 13, 23, 30, 56, 57, 10, 37]. However, a new trend is to utilize transformer blocks for this task, inspired by their success in the natural language processing domain [44]. SwinIR [29] is one such attempt that demonstrated very promising results, yet it has some fundamental downsides. One of them is that in order to work with different resolutions and upscale values, one needs to train separate networks because it uses sub-pixel convolution in the upsampling process [40]. Another is the fact that, as revealed by LAM analyses [16], it doesn't utilize long-range spatial information when it comes to activating input pixels, indicating that the memory potential of the transformer architecture is not being fully leveraged [7].

There have been two recent papers that tackled these issues. Local Texture Estimator (LTE) [26] aims to solve the first mentioned one by using an MLP to recognize the implicit representation of images and leveraging it to produce arbitrary scale super-resolution (SR) outputs, following the footsteps of LIIF [8]. It also improves upon LIIF by incorporating a Fourier Analysis component to combat a phenomenon called spectral bias [39]. This is the tendency of MLP models to favor low-frequency components of images due to their structure, and the duality of the magnitude of frequencies in the Fourier and the time domain is utilized to reduce this bias [36, 41]. On the other hand, the Hybrid Attention Transformer (HAT) [7] architecture introduces modifications to the SwinIR [29] architecture by adding a Channel Attention Block (CAB) to the original STL, and also an Overlapping Cross-Attention Block (OCAB) to the end of the residual Swin Transformer blocks (RSTB), among other things [7]. These modifications empower the network to use a hybrid attention model that utilizes both self and channel attention, which improves its ability to incorporate input pixels and better its spatial range.

We propose a modified version of LTE, which traditionally uses SwinIR as its encoder, that incorporates the mentioned advancements introduced by HAT. In doing so, we train a smaller version of LTE-SwinIR, and LTE-HAT from scratch with the DIV2K dataset from ImageNet [11], and compare the results to demonstrate that the hybrid model has better PSNR performance compared to the standalone model. Furthermore, through a combination of fine-tuned training epochs, the number of training images, and the learning rate schedule, we demonstrate that the novel small model still performs reasonably well given its size.

2. Related Work

2.1. Deep Learning in SISR

As the prominence of low-level computer vision tasks such as SISR and Image Denoising have risen in recent years, there have been many learning-based models have been developed to address the need [25, 30, 45, 54, 56, 15, 55, 28, 53, 52]. One of the most foundational papers, SRCNN [12], proposes a very straightforward convolutional neural network (CNN). This includes three convolutional layers, the first for patch extraction and representation, the second for non-linear mapping, and the third for reconstructing the HR image. Despite its simplicity, the model outperformed traditional approaches by a large margin [17, 43, 42, 35, 19] which ushered in an era dominated by other CNN-based models. As time progressed, new techniques, such as including residual blocks [23, 4, 52], dense blocks [45, 57, 58], and other modifications [9, 24, 55] have been introduced to increase the performance.

2.2. Transformers in SR: SwinIR & HAT

Despite the success of CNN-based models that incorporate these new ideas, they are held back by fundamental architectural setbacks. For example, the interactions between input images and the convolution layer are content-independent, such that CNN-based architecture might not be the best choice for SR task. Furthermore, since convolution is a naturally local operation, it lacks effectiveness in long-range modeling, despite the growing perceptive fields in deeper networks [29].

To handle these problems, Transformers [44] can come to the rescue, especially when dealing with image restoration [6, 3]. However, despite the benefits they bring to the table, they come with their own set of problems. One such problem is that, due to taking input images in patches, they may produce artificial borders in the output at places that correspond to the separation points of the patches [6, 3]. Luckily, Swin Transformer [32] shows the ability to address the problems from CNN-based models while avoiding these common issues.

SwinIR [29] is an architecture based on Swin Transformers. It consists of three parts: shallow feature extraction, deep feature extraction, and high-quality (HQ) image restoration. While the general architecture as a whole is similar to SRCNN [12], it incorporates a key difference inside the deep feature extraction section, which originally corresponds to the Non-linear mapping layer. Instead of CNN modules, it uses residual Swin Transformer Blocks, called RSTB modules. Each of these blocks is composed of several Swin Transformer Layers (STL) for local attention and cross-window interaction. As a final layer, it incorporates a convolutional layer for enhancing the final features that leave the feature extraction layer. These features are then combined with a shortcut connection from the shallow features and fused together before being upsampled.

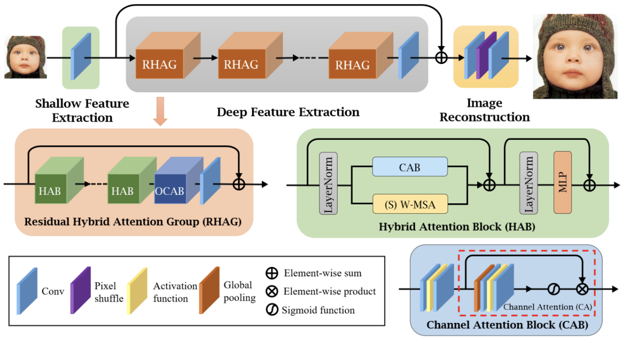

Figure 1. Architecture of HAT, demonstrating the additions of the OCAB and CAB blocks to the original SwinIR architecture [29, 7].

One thing also worth mentioning is that SwinIR contains fewer parameters compared to state-of-the-art CNN-based models [18, 56, 45, 10, 59, 37, 34, 6], which means that it requires less computational power when it comes to both training and inference. As a result, it is fair to say that Swin Transformer-based models act as a great backbone for SR.

Although the power of SwinIR is impressive compared to the other competing models, an analysis [7] which utilized the method of Local Attribution Maps (LAM) [16] found that SwinIR does not actually utilize more input pixels than CNN-based methods. Furthermore, it identified that the use of window partition mechanisms causes blocking artifacts in intermediate features, which showed that there is room for improvement in the cross-window connection of the original paper.

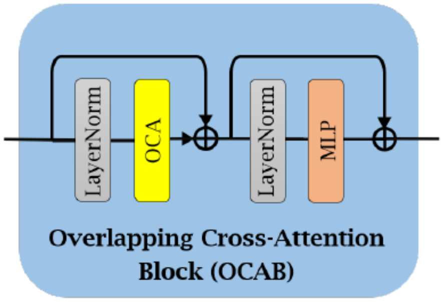

Figure 2. The structure of OCAB.

The paper that identified these issues proposes a new architecture called HAT [7], which implements two new blocks to address them. Looking at the overall architecture of HAT in Figure 1, we see that it proposes the incorporation of an overlapping cross-attention block (OCAB) after the last Hybrid Attention Block (HAB) inside the Residual

Hybrid Attention Group (RHAG). Another modification is inside the HAB block, where a Channel Attention Block (CAB) is inserted after the first layer norm.

The purpose of inserting OCAB (Figure 2) right after the last HAB block is to enlarge the receptive field for the window-based self-attention, and to better aggregate cross-window information [6, 27]. On the other hand, more pixels are activated by incorporating the CAB module inside HAB, since global information is used to calculate the channel attention weights. This further enhances the representation learning ability of this model [27, 21, 47, 38].

The PSNR performance of HAT surpasses other state-of-the-art models nowadays [30, 56, 10, 59, 37, 34, 31], and the results state in the paper show that it works particularly well for recovering images that contain high-frequency texture. This can be attributed to the fact that HAT fixes the blocking artifacts in the intermediate features, and also activates more input pixels with the use of the CAB block [7].

2.3. Arbitrary Scale SR: LIIF & LTE

A major drawback of recent approaches is that they often cannot upscale an input image by an arbitrary ratio, requiring separate models for each upscale value. To deal with this problem, LIIF [8] replaces the sub-pixel convolution that is traditionally used for upsampling the features with an MLP that uses ReLU as the activation function. The main idea is to utilize the MLP to take in continuous coordinates and latent variables as input, and then to understand the underlying implicit function. This representation can then be used to upsample the output at an arbitrary resolution.

However, it has been shown that standalone MLPs cause the model to have a bias in learning low-frequency components [39] and failing to capture fine details [41]. This phenomenon is referred to as spectral bias. Research [39] suggests that passing the input coordinates into a high-dimensional Fourier Feature space first, rather than passing them directly to LIIF might be a good idea [41], due to the spectral duality of the Fourier domain.

An LTE-based neural network [26] proposes combining these aforementioned modules to achieve arbitrary scaling while avoiding spectral bias. The overall architecture of this LTE-based model is shown in Figure 3. It uses a backbone encoder such as SwinIR [29] without the upscaling parts to extract deep features. These features are then passed into an LTE module that estimates Fourier information, reducing the spectral bias of the subsequent MLP layer in the decoder that upsamples to produce the HR output.

3. Method

The proposed model has the same fundamental architecture as the arbitrary-scale super-resolution network from the LTE paper, which can be seen from Figure 3. The model uses the encoder backbone (E_φ), which was originally the SwinIR network without its upsampling layers, to extract features from the LR image using deep transformer blocks [26]. Next, inside the pink-shaded section (LTE block), estimations on the amplitude, frequency, and phase components are made, which denote the Fourier-Domain metrics. In doing so, 3x3 convolutions layers of 256 output channels are utilized for extracting the amplitude and the frequency from the feature map given by E_φ, which are then passed through nearest-neighborhood interpolation. The phase is derived by running the cell information through a single full-connected layer with 128 hidden dimensions. Next, and inner product is carried out between the predicted frequency and the input local grid, to which the estimated phase is directly added. After passing this output through the sinusoidal activations, the result is element-wise multiplied with the predicted amplitude and then passed on to the decoder section. Finally, f_θ, which consists of a 4-Layer MLP with 256 hidden dimensions, leverages the implicit representation to upsample the results, and a bilinear-upscaled drop-connection from the raw input is added on top of this to help with the DC offset.

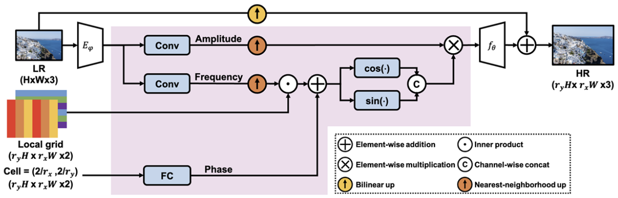

Figure 3. Overall architecture of the LTE paper. E_φ is the encoder, the pink-shaded section is the LTE, and f_θ is the decoder [26].

Our contribution has been incorporating the modifications of HAT on top of SwinIR in the encoder section, and using half as many deep feature extraction layers. To that end, we added the additional term from the CAB block, introduced the calculation of separate relative-position index (RPI) terms for the multi-head self-attention (MSA), and added the OCAB module to the end of the original RSTB block. Furthermore, we show that even by using half as many RSTB/RHAG blocks, the network is still able to produce promising results, despite the constrained training pipeline.

4. Experiments

We use a portion of the DIV2K [1] dataset for training the network, which consists of 600 2K-resolution images. For evaluating our network, we use Set14 [51], Urban100 [20], and BSD100 [33].

Since our model is directly based on the work of the LTE paper, we used many of the same choices when it comes to implementing the model. These include the use of 48 x 48 input patches, utilizing bicubic down-sampling for training, leveraging L1 distance for the loss function, and using

Adam as the optimized for training the model.

Where we diverge has to do with hyperparameters that relate to the training pipeline. Due to the hardware constraints and limitations, we had to downscale the original encoder significantly, and we conducted tests to see the best areas of compromise. The metrics that determined the size of the encoder architecture included the depth of HAT/SwinIR, and the number of RHAG/RSTB and HAB/STL blocks used. These are parametrized and provided to the network as arguments, and by default, they are all 6. This is represented with setting both the depths and the num_heads arguments to [6, 6, 6, 6, 6, 6] [7].

We also tried to reduce the number of training images and the number of epochs as much as possible without excessively compromising the model. Furthermore, due to the structural change, we looked into better selecting the learning rate, and its schedule. The training loss over epochs plots shown in Figure 4 demonstrates our trials over various models under different settings. The labels of each plot indicate various attributes, where the structure is learning-rate_model-size_number-of-images_model. The model size is denoted by a multiplication, where the first number is the number of repeats of the second number in the list provided as arguments to SwinIR/HAT. For example, 4x6 denotes using depts=[6, 6, 6, 6], whereas 6x2 denotes [2, 2, 2, 2, 2, 2].

These results gave us many insights. First off, even though generative models are not as prone to overfitting, we realized that using more images tended to make the training process more stable. This can be inferred by analyzing that the loss function decrease is smoother over epochs for the training pipelines that used more images. Secondly, we saw that reducing the depth of the HAT/SwinIR modules, rather than the number of interior blocks, provided a better return, which can be seen from comparing the models that have 2x6 and 6x2 as their architectures. We saw that the original learning rate performed well enough when compared to the alternatives of 5e-5 and 1e-3, and in providing a healthy curve. Lastly, as is evidenced by the training of 2e-4_2x6_600_LTE-SwinIR, the use of 600 training images proves to be a happy medium for decreasing the number of images without compromising the stability, but 2x6 is too small of a model size for extended training sessions.

In light of these results, we settled on using 2e-4 for the learning rate, trained our network for 350 epochs, and a more aggressive learning schedule comprising of milestones at 200, 300 and 330. In terms of model size, we settled on 3x6 (setting depths and num_heads to [6, 6, 6]).

We trained both the hybrid (LTE-HAT), and the standalone baseline (LTE-SwinIR) with these parameters, labeled as LTE-HAT_Final and LTE-SwinIR_Final respectively, and it can be seen that their losses show a healthy progression during training. In terms of time, it took our setup, which had a laptop-grade Nvidia RTX 3070, 5 hours to train the hybrid, and 3.2 hours to train the baseline model.

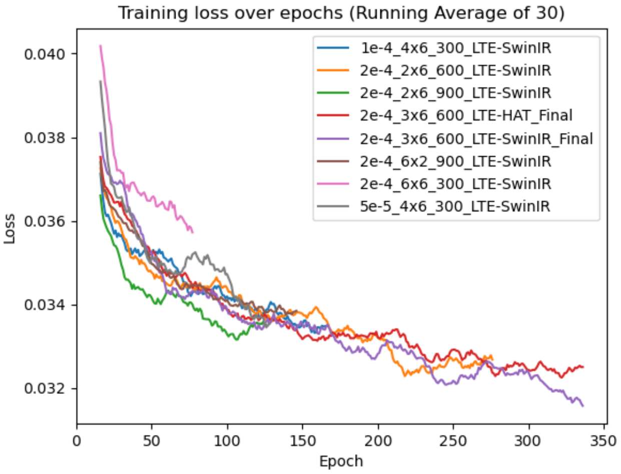

Figure 4. Training loss plots for various model configurations labeled as **learning-rate_model-size_number-of-images_model**.

PSNR Performance (dB) Over Evaluation Datasets

Model (Small Version)	SET14	Urban100	BSD100
LTE-SwinIR (Baseline)	31.4325	29.5829	30.6286
LTE-HAT (Ours)	**31.4894**	**29.6883**	**30.6588**

Table 1. Evaluation comparison of baseline and proposed model

The final results displaying the PSNR values of both models on the Set14, Urban100, and BSD100 evaluation datasets (all x2 bicubic) can be seen in Table 1, which shows that our proposed model (LTE-HAT) has better performance compared to the baseline (LTE-SwinIR) for all tested datasets, and that it achieves respectable effectiveness despite the constraints.

5. Conclusion

In this paper, we propose incorporating modificaitons of HAT to the SwinIR model, which is then used as the encoder of the arbitrary scale SR architecture proposed by LTE. Our model leverages the strengths of using MLPs for operating at arbitrary scale with a single model, using the Fourier Analysis to combat spectral bias, and incorporating the CAB, OCAB, and other advancements from HAT in a single model. Comparisons with the baseline show that our proposed solutions provide a meaningful increase in PSNR performance when tested on the Set14, Urban100, and BSD100 evaluation datasets. The biggest weakness is the size of the model and the constrained training regiment as a result of hardware and time constraints. Even though this simple model shows promising performance, future work is desirable to ensure that these performance gains persist in a full-sized model as well. To reproduce our work, please use the provided GitHub repository.

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. 3

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 1

[3] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021. 2

[4] Lukas Cavigelli, Pascal Hager, and Luca Benini. Cas-cnn: A deep convolutional neural network for image compression artifact suppression. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 752–759. IEEE, 2017. 2

[5] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004. 1

[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. 2, 3

[7] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. *arXiv preprint arXiv:2205.04437*, 2022. 1, 2, 3, 4

[8] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 1, 3

[9] 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. 2

[10] 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/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019. 1, 2, 3

[11] 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

[12] 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

[13] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016. 1

[14] William T Freeman, Egon C Pasztor, and Owen T Carmichael. Learning low-level vision. *International journal of computer vision*, 40(1):25–47, 2000. 1

[15] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3599–3608. IEEE, 2019. 2

[16] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021. 1, 2

[17] Shuhang Gu, Nong Sang, and Fan Ma. Fast image super resolution via local regression. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 3128–3131. IEEE, 2012. 2

[18] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673, 2018. 2

[19] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010. 2

[20] 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. 3

[21] Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, and Bin Fu. Shuffle transformer: Rethinking spatial shuffle for vision transformer. *arXiv preprint arXiv:2106.03650*, 2021. 3

[22] Kui Jia, Xiaogang Wang, and Xiaoou Tang. Image transformation based on learning dictionaries across image spaces. *IEEE transactions on pattern analysis and machine intelligence*, 35(2):367–380, 2012. 1

[23] 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. 1, 2

[24] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 2

[25] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1, 2

[26] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1929–1938, 2022. 1, 3

[27] Wenbo Li, Xin Lu, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 3

[28] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3867–3876, 2019. 2

[29] 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. 1, 2, 3

[30] 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. 1, 2, 3

[31] Zudi Lin, Prateek Garg, Atmadeep Banerjee, Salma Abdel Magid, Deqing Sun, Yulun Zhang, Luc Van Gool, Donglai Wei, and Hanspeter Pfister. Revisiting rcan: Improved training for image super-resolution. *arXiv preprint arXiv:2201.11279*, 2022. 3

[32] 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. 2

[33] 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. 3

[34] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3517–3526, 2021. 2, 3

[35] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 945–952, 2013. 2

[36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[37] 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. 1, 2, 3

[38] Krushi Patel, Andres M Bur, Fengjun Li, and Guanghui Wang. Aggregating global features into local vision transformer. *arXiv preprint arXiv:2201.12903*, 2022. 3

[39] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019. 1, 3

[40] 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. 1

[41] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 1, 3

[42] R Timofte, V De Smet, and LA Van Gool. Adjusted anchored neighborhood regression for fast super-resolution. inasian conference on computer vision 2014 nov 1 (pp. 111-126). 2

[43] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 1920–1927, 2013. 1, 2

[44] 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. 1, 2

[45] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 2

[46] Alexander Watson. Deep learning techniques for super-resolution in video games. *arXiv preprint arXiv:2012.09810*, 2020. 1

[47] Sitong Wu, Tianyi Wu, Haoru Tan, and Guodong Guo. Pale transformer: A general vision transformer backbone with pale-shaped attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2731–2739, 2022. 3

[48] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang. Coupled dictionary training for image super-resolution. *IEEE transactions on image processing*, 21(8):3467–3478, 2012. 1

[49] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 1

[50] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010. 1

[51] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010. 1, 3

[52] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[53] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind

image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 2

[54] 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. 2

[55] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3262–3271, 2018. 2

[56] 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. 1, 2, 3

[57] 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. 1, 2

[58] 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*, 43(7):2480–2495, 2020. 2

[59] 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*, 33:3499–3509, 2020. 2, 3

[60] Wilman WW Zou and Pong C Yuen. Very low resolution face recognition problem. *IEEE Transactions on image processing*, 21(1):327–340, 2011. 1