Channel-Partitioned Windowed Attention And Frequency Learning for Single Image Super-Resolution

Supplementary Material

Dinh Phu Tran
phutx2000@kaist.ac.kr
Dao Duy Hung
hicehehe@kaist.ac.kr
Daeyoung Kim
kimd@kaist.ac.kr

Korea Advanced Institute of Science and Technology
Daejeon, Korea

1 Experimental Setting Details

1.1 Training Details

We use DF2K (DIV2K [1]+Flicker2K [11]) with 3450 high-resolution images as the training dataset. The low-resolution (LR) images are generated from the ground truth images by the BICUBIC downsampling in MATLAB. We evaluate experimental results with PSNR and SSIM [13] values on Y channel of images transformed to YCbCr space. We use the input patch size is 64×64, and the mini batch size is 32 with total training iterations are set to 500K. The learning rate is initialized as $2e-4$ and reduced by half at [250K, 400K, 450K, 475K], where 1K means one thousand. For data augmentation on training data, we use geometric data augmentations are random rotation of 90°, 180°, 270° and horizontally flipping. For x4 SR , we utilize the pre-trained model x2 SR weight and halve the iterations for each learning rate decay as well as total iterations. We simply use L1 loss function, Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and zero weight decay to train our model. Our CPAT is implemented on PyTorch [15] framework with 8 NVIDIA RTX6000 GPUs.

1.2 Network Structure Details

For the base model, we set the number of RWAG and SPWin-SA to 6, the channel number to 180, and the window size is set to 16. The overlapping ratio in OCAM is set to 0.5. For the lightweight model, we keep the settings the same as the base model except for the following settings. We reduce the channel number from 180 to 51. We use 4 RWAG modules, and the number of SPWin-SA is also reduced to 4, except for the last block of RWAG, where we use 5 SPWin-SA modules. When applying the self-ensemble strategy [16], "†" is added after the model name called CPAT†, and we only apply for base CPAT model.

2 Self-ensemble Strategy

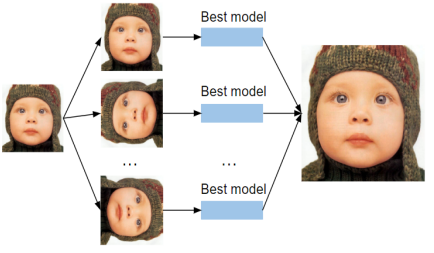

Figure 1: Self-ensemble strategy.

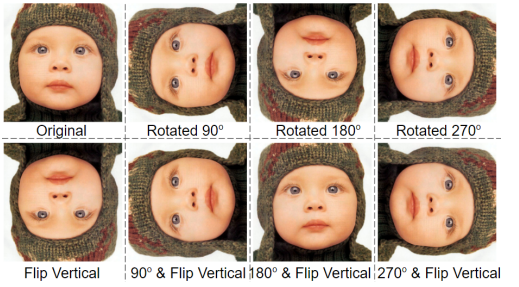

Figure 2: Augmentation of test images for self-ensemble strategy by rotation and flip.

We apply self-ensemble to our network during the testing phase. For each input image I^{LR} in testing time, we perform transformations: rotate 90°, rotate 180°, rotate 270°, rotate 90° + flip vertical, rotate 180° + flip vertical, and rotate 270° + flip vertical, resulting in a total of eight images $I_i^{LR} = T_i(I_n^{LR})$ including the original one, where T_i represents the i-th transformation, $i = 1, 2, ..., 8$ including identity, which are shown in Fig. 2. We then use the best model to generate high-resolution images $\{O_1^{SR}, ..., O_8^{SR}\}$ for these transformed images which was described in Fig. 1. We then apply inverse transformation to those SR images to get the original geometry $\hat{O}_i^{SR} = T_i^{-1}(O_i^{SR})$, where T_i^{-1} represents the i-th inverse transformation. Finally, we simply average these outputs to obtain the final result $O_{SR} = \frac{1}{8}\sum_{i=1}^{8} \hat{O}_i^{SR}$. Although this method can improve the performance of our model, the inference time for each image is significantly increased because we need to run with eight images instead of one image. Therefore, the self-ensemble strategy is not effective when applied in practice, especially with high-resolution images.

3 Extensive Experiments

3.1 Quantitative Results of Lightweight Model

We compare our lightweight model (CPAT-light) to the state-of-the-art lightweight methods. In addition to PSNR/SSIM [18], we compare model size and computational complexity. We report the number parameter for model size comparison and multiply-accumulate operations (evaluated on a 1280x720 HR image) for a fair computational complexity comparison. Tab. 1 shows that CPAT-light outperforms the state-of-the-art methods. Specifically, CPAT-light surpasses the current state-of-the-art method of lightweight SISR Omni-SR by up to 0.31dB on Urban100, which indicates our method is efficient and able to work well on different scales of model size. By enhancing the window size along the width and height of the input feature maps in V-EWin and H-EWin, our method can extract more global contextual information and relationships between the distant tokens, thereby improving performance of model. CPAT-light outperforms current methods in terms of PSNR/SSIM metrics and maintains a model size and computational complexity similar to other methods. Specifically, the number of parameters of CPAT-light is always less than 1M (939K) parameters for all scales, while the #Mult-Adds is comparable to other methods. These results demonstrate that our

method can perform well even with a lightweight version. These results show that enhancing window size still efficient for our lightweight model. Furthermore, leveraging frequency features makes our method more robust.

Method	Scale	#Params	#Mult-Adds	Set5 [0]		Set14 [10]		BSD100 [6]		Urban100 [8]		Manga109 [9]	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LAPAR-A [0]	x2	548K	171.0G	38.01	0.9605	33.62	0.9183	32.19	0.8999	32.10	0.9283	38.67	0.9772
SCET [22]		683K	70.5G	38.06	0.9615	33.78	0.9198	32.24	0.9006	32.38	0.9299	39.86	0.9821
FMEN [0]		748K	172.0G	38.10	0.9609	33.75	0.9192	32.26	0.9007	32.41	0.9311	38.95	0.9778
ASSLN [21]		692K	159.1G	38.12	0.9608	33.77	0.9194	32.27	0.9007	32.41	0.9309	39.12	0.9781
LKASR [0]		947K	141.0G	38.25	0.9614	34.17	0.9228	32.39	0.9023	33.10	0.9375	39.50	0.9786
Omni-SR [23]		772K	-	38.29	0.9617	34.27	0.9238	32.41	0.9026	33.30	0.9386	39.53	0.9792
CPAT-light (Ours)		939K	143.3G	38.36	0.9618	34.31	0.9243	32.45	0.9031	33.62	0.9405	39.72	0.9795
LAPAR-A [0]	x3	594K	114G	34.36	0.9267	30.34	0.8421	29.11	0.8054	28.15	0.8523	33.51	0.9441
SCET [22]		683K	70.5G	34.53	0.9278	30.43	0.8441	29.17	0.8075	28.38	0.8559	34.29	0.9503
FMEN [0]		757K	77.2G	34.45	0.9275	30.40	0.8435	29.17	0.8063	28.33	0.8562	33.86	0.9462
ASSLN [21]		698K	71.2G	34.51	0.9280	30.45	0.8439	29.19	0.8069	28.35	0.8562	34.00	0.9468
LKASR [0]		947K	70.5G	34.74	0.9296	30.66	0.8481	29.30	0.8098	28.93	0.8674	34.45	0.9496
Omni-SR [23]		780K	-	34.77	0.9304	30.70	0.8489	29.33	0.8111	29.12	0.8712	34.64	0.9507
CPAT-light (Ours)		939K	60.6G	34.80	0.9306	30.73	0.8495	29.36	0.8122	29.39	0.8751	34.76	0.9516
LAPAR-A [0]	x4	659K	94.0G	32.15	0.8944	28.61	0.7818	27.61	0.7366	26.14	0.7871	30.42	0.9074
SCET [22]		683K	70.5G	32.27	0.8963	28.72	0.7847	27.67	0.7390	26.33	0.7915	31.10	0.9155
FMEN [0]		769K	44.2G	32.24	0.8955	28.70	0.7839	27.63	0.7379	26.28	0.7908	30.70	0.9107
ASSLN [21]		708K	40.6G	32.29	0.8964	28.69	0.7844	27.66	0.7384	26.27	0.7907	30.84	0.9119
LKASR [0]		1026K	62.6G	32.63	0.8998	28.94	0.7894	27.78	0.7430	26.79	0.8068	31.42	0.9187
Omni-SR [23]		792K	-	32.57	0.8993	28.95	0.7898	27.81	0.7439	26.95	0.8105	31.50	0.9192
CPAT-light (Ours)		939K	33.9G	32.65	0.9009	28.97	0.7905	27.83	0.7454	27.10	0.8152	31.60	0.9215

Table 1: Quantitative comparison with state-of-the-art lightweight methods. The best and second-best results are marked in red and blue colors, respectively.

3.2 Extensive Qualitative Results for Base Model

Fig. 3 and Fig. 5 show more qualitative results for our base model (CPAT) and comparison with HAT. In the Fig. 3, for all images, our method provide better results both of LAM attribution visualization and DI values. LAM visualizations and DI values show that our method use more and wide range pixel during upscaling the patches that marked on the green boxes. Fig. 5 demonstrates that our method can reconstruct HR images more clearly and retain the SR image details better than HAT. Particularly, SR images of "*img*_011" and "*img*_060" in Urban100 generated by HAT can not keep some details of HR images whereas our method can do that much better.

3.3 Extensive Ablation Study

Following [2, 5], we train x2 SR the model on DF2K (DIV2K [1]+Flicker2K [11]), and test on Urban100 [8] for all experiments in this section. FLOPs is calculated on a 256x256 HR image. Results are reported in the Tab. 2, 3 and the better results are shown in **bold**.

Effective Receptive Field Analysis. For image super-resolution, enlarging receptive field extract the global context information is crucial for HR image reconstruction. We use effective receptive field (ERF) as a toolkit to visualize the effectiveness of our proposed CPAT compared to other SOTA methods. In Fig. 4, we visualize ERFs of RCAN, SwinIR, HAT and CPAT on Urban100. We can make the following observations: 1) For CNN-based method, the effective receptive field is limited (local ERF). By contrast, Transformer-based methods can reach global ERFs. 2) Our CPAT is the only model that can achieve a significant global effective receptive field, thus improving our model's performance in SISR.

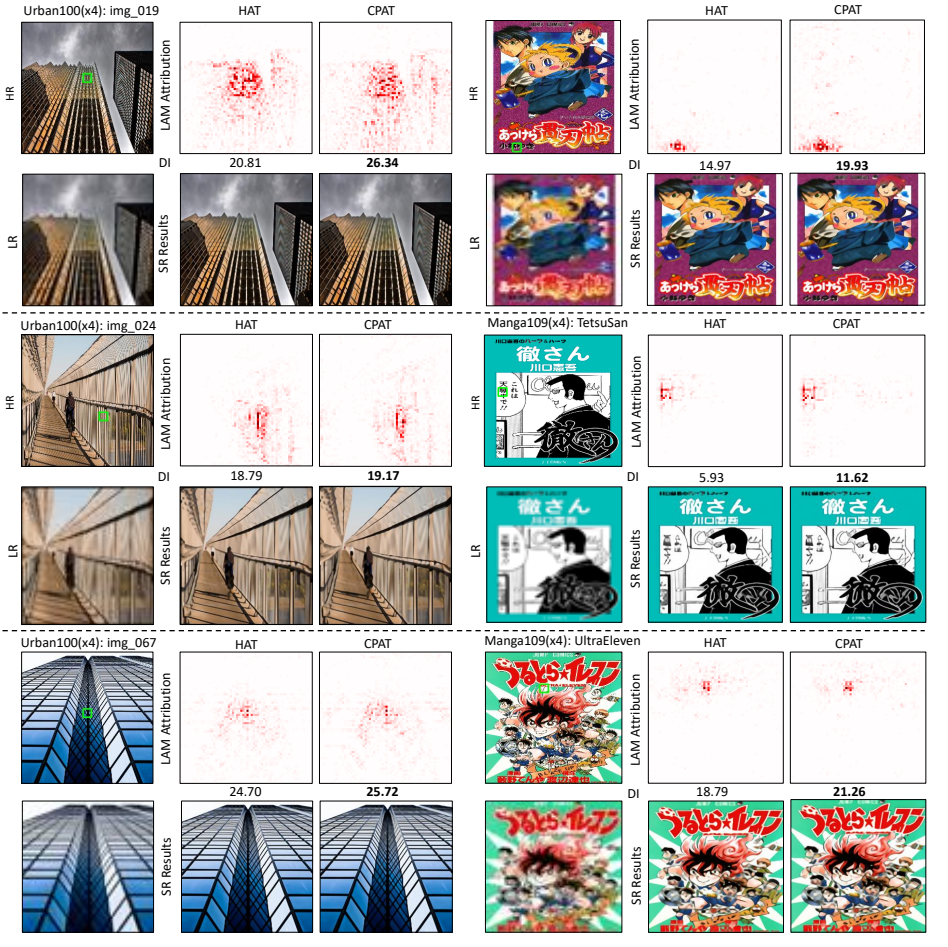

Figure 3: Comparison of LAM between CPAT and HAT.

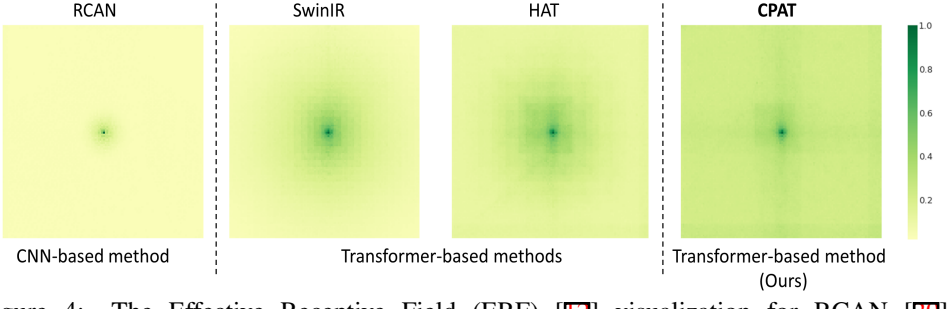

Figure 4: The Effective Receptive Field (ERF) [12] visualization for RCAN [21], SwinIR [10], HAT [3] and our proposed CPAT on Urban100. A larger ERF is indicated by a more extensively distributed dark area. Only the proposed CPAT achieves a significant global effective receptive field.

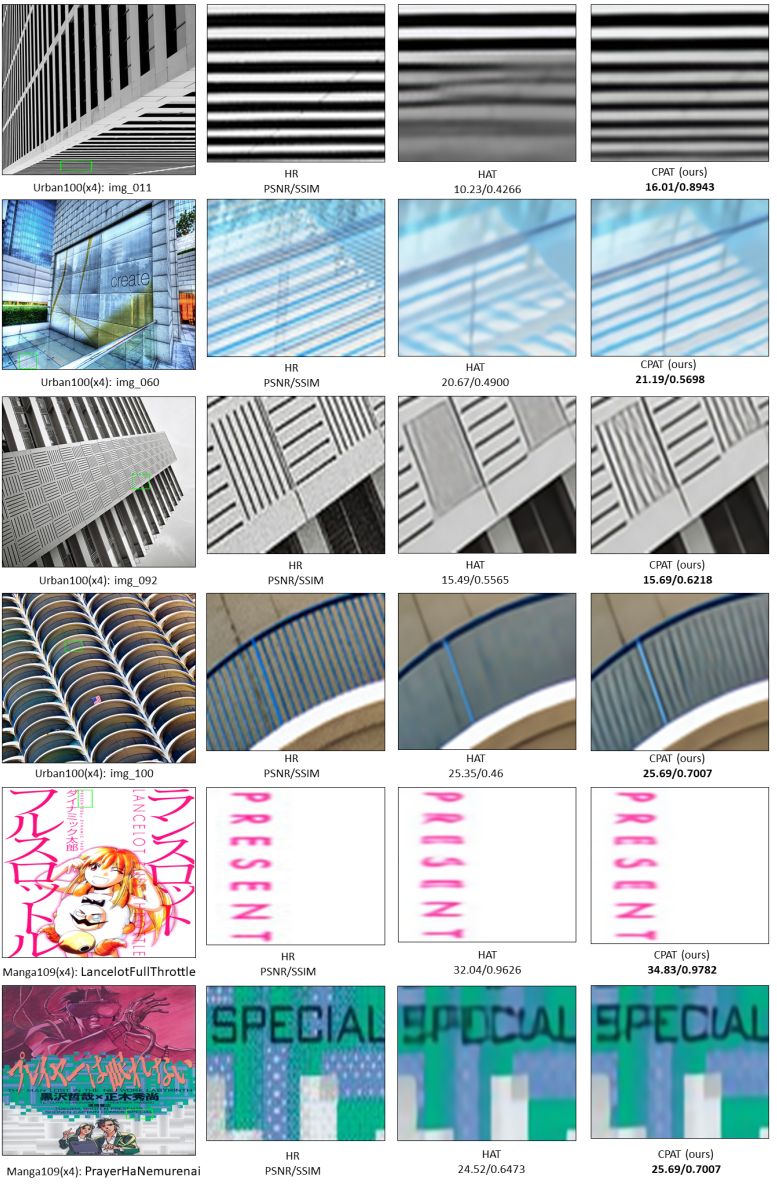

Figure 5: Visual comparisons between CPAT and HAT.

#Channels	90	120	180	210
PSNR	33.64	33.95	34.26	**34.35**
SSIM	0.9404	0.9428	0.9448	**0.9456**
FLOPs	90.03G	152.22G	329.04G	443.66G
#Params	5.33M	9.25M	20.39M	27.61M

Table 2: Effect of number channels.

Effect of number channel. Effect of number channel on the performance's the models is shown in Tab. 2. If we gradually increase the number of channels, the model's performance will continue to improve. This result shows that our model is scalable and can perform well if we downscale the model size. The biggest model (210 channels) reaches 34.35dB, whereas the smallest one (90 channels) gets 33.64dB. However, it is worth noting that when the number channel becomes too large, the improvement in the model's performance does not increase

significantly compared to the training cost required. Therefore, based on the performance as well as the complexity of our model, we set the number channel for the base model to be 180. This also helps us to compare fairly with state-of-the-art methods that also use 180 channels in these methods.

Effect of SFIM variants. We propose three potential designs, shown in Fig. 6 of SFIM, from which we select SFIM-1 as the optimal design, which reaches 34.26dB in PSNR. It is similar to SFIM-3 in terms of quantitative results, but simpler and has the lowest floating-point operations per second (FLOPs) and number of parameters. Tab. 3 shows that SFIM-1 is an optimal design based on performance and complexity trade-offs, which we choose for our base and lightweight models.

Structure	PSNR	SSIM	#Params	FLOPs
w/ SFIM-1	**34.26**	0.94484	20.39M	329.0G
w/ SFIM-2	34.20	0.94433	21.13M	330.9G
w/ SFIM-3	**34.26**	**0.94486**	20.96M	332.8G

Table 3: Effect of SFIM variants

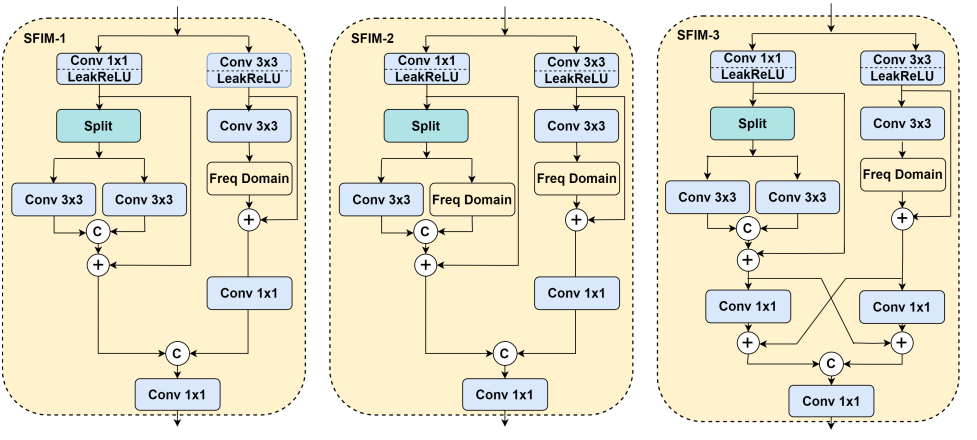

Figure 6: Three potential variants of SFIM. SFIM-1 is chosen for use in the base model.

References

[1] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1122–1131, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. doi: 10.1109/CVPRW.2017.150. URL https://doi.ieeecomputersociety.org/10.1109/CVPRW.2017.150.

[2] Marco Bevilacqua, Aline Roumy, Christine M. Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *British Machine Vision Conference*, 2012. URL https://api.semanticscholar.org/CorpusID:5250573.

[3] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22367–22377, 2023.

[4] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Cross aggregation transformer for image restoration. *Advances in Neural Information Processing Systems*, 35:25478–25490, 2022.

[5] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12312–12321, 2023.

[6] Zongcai Du, Ding Liu, Jie Liu, Jie Tang, Gangshan Wu, and Lean Fu. Fast and memory-efficient network towards efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 853–862, 2022.

[7] Hao Feng, Liejun Wang, Yongming Li, and Anyu Du. Lkasr: Large kernel attention for lightweight image super-resolution. *Know.-Based Syst.*, 252(C), sep 2022. ISSN 0950-7051. doi: 10.1016/j.knosys.2022.109376. URL https://doi.org/10.1016/j.knosys.2022.109376.

[8] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015. doi: 10.1109/CVPR.2015.7299156.

[9] Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, and Jiaya Jia. Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. *Advances in Neural Information Processing Systems*, 33: 20343–20355, 2020.

[10] 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.

[11] 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.

[12] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

[13] 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*, volume 2, pages 416–423 vol.2, 2001. doi: 10.1109/ICCV.2001.937655.

[14] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, November 2016. ISSN 1573-7721. doi: 10.1007/s11042-016-4020-z. URL http://dx.doi.org/10.1007/s11042-016-4020-z.

[15] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[16] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1865–1873, 2016.

[17] Hang Wang, Xuanhong Chen, Bingbing Ni, Yutian Liu, and Jinfan Liu. Omni aggregation networks for lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22387, 2023.

[18] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[19] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. volume 6920, pages 711–730, 06 2010. ISBN 978-3-642-27412-1. doi: 10.1007/978-3-642-27413-8_47.

[20] 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.

[21] Yulun Zhang, Huan Wang, Can Qin, and Yun Fu. Aligned structured sparsity learning for efficient image super-resolution. In *NeurIPS*, 2021.

[22] Wenbin Zou, Tian Ye, Weixin Zheng, Yunchen Zhang, Liang Chen, and Yi Wu. Self-calibrated efficient transformer for lightweight super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 930–939, 2022.