Misalignment-Robust Frequency Distribution Loss for Image Transformation
Supplementary Material

Zhangkai Ni[1], Juncheng Wu[1†], Zian Wang[1†], Wenhan Yang[2*], Hanli Wang[1*], Lin Ma[3]
[1]Tongji University, [2]Peng Cheng Laboratory, [3]Meituan

This supplementary material serves as an appendix to our main paper. In Section 1, we provide more details of the experiments conducted in the paper. In Section 2, we analyze the specific misalignment phenomena present in our synthetic DIV2K dataset and real-world datasets. Section 3 presents additional experimental results that complement the main findings. Lastly, Section 4 showcases more visual results for various image transformation tasks with misalignment.

1. Experiment Details

In this section, we present additional experiment details that are not included in the main paper due to space limitations.

Implementation details of FDL. During the computation of the proposed Frequency Distribution Loss (FDL), features from the $Relu\_1\_1$, $Relu\_2\_1$, $Relu\_3\_1$, $Relu\_4\_1$, and $Relu\_5\_1$ layers of VGG19 [20] are utilized. The average of the distances calculated in these different layers is taken as the final result of the FDL. Sliced Wasserstein Distance [9] is employed to measure the distance between distributions, utilizing a set of random linear projections. The number of projections is set to 256 in our work.

Shift Response Curves. We aim to investigate the robustness of the proposed loss to geometric misalignment by plotting the shift response curves. Specifically, for a given image I, a series of patches P_i are cropped. This process can be formulated as:

$$P_i = I[0:p, i:(i+p)], i \in [0, n], \tag{1}$$

where p represents the patch size and n represents the quantity of patches. To visualize the shift response of different loss functions, we calculate the normalized loss between image P_i and P_0:

$$\mathcal{R}_M^i = \frac{\mathcal{L}_M(P_0, P_i)}{\mathcal{L}_M(P_0, P_0+\epsilon)}, \tag{2}$$

[†]Equal contribution. [*]Corresponding author.

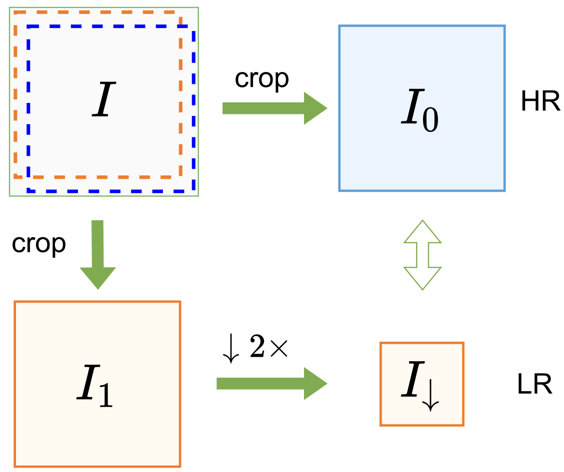


Figure 1. Process of creating a dataset with random misalignment based on the DIV2K [1].

where $\mathcal{L}_M$ represents different loss functions. The denominator term is utilized to minimize the impact of varying magnitudes of different loss functions, where ϵ represents the Gaussian noise:

$$\epsilon \sim \mathcal{N}(0, \sigma_0), \tag{3}$$

where σ_0 equals to the standard deviation of P_0. Through calculating the distance between the distorted image $P_0+\epsilon$ and the original one P_0, the results $\mathcal{R}_M$ can reflect the relative significance of the responses of $\mathcal{L}_M$ caused by misalignment compared to noise degradation. In this work, n is set to 40, and p is set to 256. Two classical element-wise loss functions MSE and LPIPS [22] are adopted for comparison. To mitigate the influence of the content of images, the average $\mathcal{R}_M^i$ of 100 different I from DIV2K[1] is used for plotting.

2. Misalignment in Datasets

Synthetic DIV2K Dataset. Based on the DIV2K [1] dataset, we synthesize a single image super-resolution dataset with significant misalignment, as illustrated in Figure 1. For each image I in DIV2K, we randomly crop two different regions with the same resolution, resulting in two sub-images, I_0 and I_1. Then, I_1 is downsampled to generate the corresponding low-resolution (LR) image, denoted as $I_\downarrow$. This $I_\downarrow$ is paired with I_0 to create a training pair with misalignment. The two different regions used for cropping may have random misalignment within a range of 0-24 pixels. This method allows us to synthesize geometric misalignment with an unknown direction and magnitude, simulating real-world scenarios.

Real-World Single Image Super-Resolution Dataset. To validate the effectiveness of the proposed FDL for real-world data, we merge the RealSR [3] and City100 [5] datasets as the training and testing sets for real-world super resolution tasks. Both datasets consist of images captured with different camera focal lengths within the same scene, which serve as ground truth and low-resolution images, respectively. To address the misalignment between the ground truth and low-resolution images, sophisticated alignment algorithms are employed during dataset creation, resulting in nearly imperceptible misalignment in training pairs. However, the complexity of these alignment algorithms has posed challenges in dataset creation, limiting the applicability of image transformation algorithms in real-world scenarios. Furthermore, our experiments reveal that, even in the absence of significant misalignment in the training data, our proposed Frequency Distribution Loss (FDL) consistently leads to better-quality predicted results compared to other perceptual loss functions.

DPED Dataset. The DPED dataset [13] consists of images captured using DSLR cameras and mobile phones in the same scene, serving as the ground truth and low-quality images, respectively. Consequently, this inevitably introduces noticeable misalignment between the training pairs. To mitigate this misalignment, the DPED dataset employs an alignment algorithm during the preprocessing of image pairs. Specifically, the alignment algorithm in the DPED dataset utilizes SIFT [17] descriptor matching to identify corresponding regions in both the Ground Truth (GT) and Low-Quality (LQ) images. Non-linear transformations [21] and cropping are then applied to align the images. Finally, the aligned images are cropped into patches after the alignment process. Despite undergoing complex processing, we observe that the DPED dataset still has noticeable geometric misalignment, as shown in Figure 2.

3. Additional Experiment

In our main paper, we provide experimental results with misaligned scenarios by using *only* the proposed FDL and comparison loss functions in the feature domain (*i.e.*, PDL [7], LPIPS [22], CTX [18]). Previous works demonstrated that incorporating both feature-based loss and pixel loss as training constraints for the model can better balance the perceptual quality and detail fidelity of predicted images [7, 11, 14]. In this section, we jointly use various feature domain loss functions with pixel loss

(a) GT (b) LQ

Figure 2. An example of noticeable geometric misalignment within image pairs from the DPED dataset.

$$\mathcal{L}(x, y) = \mathcal{L}_1(x, y) + \alpha \cdot \mathcal{L}_f(x, y), \quad (4)$$

where x and y represent the model's predicted image and ground truth image respectively, $\mathcal{L}_1(\cdot,\cdot)$ is L1 norm between x and y, and $\mathcal{L}_f(\cdot,\cdot)$ represents various feature domain loss function, α is the weight of $\mathcal{L}_f(\cdot,\cdot)$. In all experiments, the weights of feature domain loss functions are empirically set to 0.1 and 0.01, respectively.

Results on the Synthetic DIV2K Dataset. We adopt the NLSN [19] as the baseline model and train it on our synthetic DIV2K dataset. Due to the presence of significant misalignment in this dataset, utilizing L1 loss solely as the model's training constraint leads to noticeable regression to the mean phenomenon, as shown in Figure 3. The quantitative results are detailed in Table 1, and we can observe that using L1 as an independent loss function performs poorly on all metrics. However, combining L1 with CTX, PDL, and FDL as loss functions improves the performances on all metrics, indicating that these three loss functions possess some degree of misalignment robustness. Furthermore, combining L1 with FDL consistently outperforms CTX and

(a) GT (b) L1 (c) FDL (Ours)

Figure 3. Qualitative comparison results of NLSN [19] on our synthetic shifted DIV2K dataset. Training a model using L1 loss results in regression to the mean phenomenon. In contrast, FDL exhibits misalignment robustness, ensuring the overall quality of predicted images in misaligned scenarios.

Test Set	Loss	PSNR↑	LPIPS↓	DISTS↓	SSIM↑	FID↓
Set5	L1	23.066	0.450	0.328	0.811	57.581
	LPIPS(alex) (0.1)	21.373	0.277	0.228	0.817	40.928
	LPIPS(alex) (0.01)	19.666	0.431	0.333	0.759	172.385
	CTX (0.1)	27.297	0.099	0.097	0.918	5.188
	CTX (0.01)	22.914	0.120	0.144	0.862	13.460
	PDL (0.1)	27.653	0.178	0.146	0.852	14.175
	PDL(0.01)	26.804	0.181	0.149	0.875	13.931
	FDL (0.1)	**29.019**	**0.107**	**0.098**	**0.925**	**4.009**
	FDL (0.01)	**32.154**	**0.095**	**0.095**	**0.957**	**4.450**
Set14	L1	22.286	0.539	0.352	0.749	106.836
	LPIPS(alex) (0.1)	20.905	0.329	0.237	0.782	51.432
	LPIPS(alex) (0.01)	19.396	0.494	0.360	0.704	232.432
	CTX (0.1)	25.814	0.159	0.112	0.913	7.235
	CTX (0.01)	22.327	0.176	0.153	0.854	29.962
	PDL (0.1)	26.279	0.236	0.163	0.826	19.854
	PDL(0.01)	25.479	0.236	0.165	0.858	14.472
	FDL (0.1)	**26.948**	**0.168**	**0.109**	**0.922**	**8.275**
	FDL (0.01)	**29.383**	**0.157**	**0.108**	**0.965**	**8.973**
B100	L1	23.110	0.549	0.378	0.725	229.265
	LPIPS(alex) (0.1)	21.837	0.329	0.250	0.753	175.323
	LPIPS(alex) (0.01)	20.583	0.491	0.378	0.686	389.971
	CTX (0.1)	26.206	0.156	0.118	0.868	20.851
	CTX (0.01)	23.397	0.171	0.150	0.825	48.058
	PDL (0.1)	26.554	0.227	0.173	0.800	34.974
	PDL(0.01)	25.897	0.230	0.176	0.825	41.311
	FDL (0.1)	**26.915**	**0.167**	**0.117**	**0.868**	**18.188**
	FDL (0.01)	**28.791**	**0.155**	**0.116**	**0.908**	**19.663**
Urban100	L1	19.962	0.507	0.355	0.639	92.895
	LPIPS(alex) (0.1)	19.339	0.305	0.235	0.673	56.614
	LPIPS(alex) (0.01)	17.888	0.449	0.345	0.585	199.388
	CTX (0.1)	23.469	0.149	0.111	0.815	12.744
	CTX (0.01)	20.528	0.160	0.150	0.754	46.750
	PDL (0.1)	23.750	0.225	0.163	0.716	24.561
	PDL(0.01)	23.148	0.228	0.166	0.749	21.997
	FDL (0.1)	**24.260**	**0.154**	**0.105**	**0.836**	**8.870**
	FDL (0.01)	**26.658**	**0.139**	**0.102**	**0.894**	**8.862**

Table 1. Quantitative comparison of SISR using the NLSN model [19] trained with synthetic misaligned DIV2K dataset. Each loss is combined with L1 loss using corresponding weights. The best results are marked in red.

Loss	PSNR↑	LPIPS↓	DISTS↓	SSIM↑	FID↓
L1	37.368	0.111	0.124	0.978	32.246
LPIPS(alex) (0.1)	36.954	0.096	0.105	0.976	33.766
LPIPS(alex) (0.01)	37.354	0.106	0.121	0.978	40.976
CTX (0.1)	35.907	0.097	0.105	0.969	48.428
CTX (0.01)	36.104	0.102	0.105	0.970	49.567
PDL (0.1)	34.523	0.099	0.105	0.951	51.348
PDL (0.01)	36.094	0.090	0.097	0.968	28.991
FDL (0.1)	**35.981**	**0.089**	**0.091**	**0.967**	**44.501**
FDL (0.01)	**36.215**	**0.092**	**0.096**	**0.970**	**15.172**

Table 2. Quantitative comparison of real-world SISR using the SwinIR [16] model trained with merged real-world dataset.

PDL across almost all metrics. This implies that FDL is more effective in ensuring the overall quality of predicted results in the presence of significant misalignment compared to other loss functions.

Results on the Merged Real-world Dataset We utilize SwinIR [15] as our baseline model and trained it on the merged real-world dataset. Quantitative results are presented in Table 2. It is evident that the combination of L1 and FDL as the loss function yields the best performance in terms of perceptual quality metrics, including LPIPS [22], DISTS [8], and FID [12]. The absence of significant misalignment in the dataset, coupled with the perceptual-distortion tradeoff [2], makes L1 alone the optimal loss function for preserving detail fidelity in the predicted results. This conclusion is supported by its superior performance in PSNR and SSIM metrics.

Results on the DPED Dataset. We employ SwinIR [15] as our baseline model and train it on the DPED dataset. The quantitative results are presented in Table 3. Due to the presence of noticeable misalignment in the DPED dataset, using L1 as the sole loss function exhibits mediocre perfor-

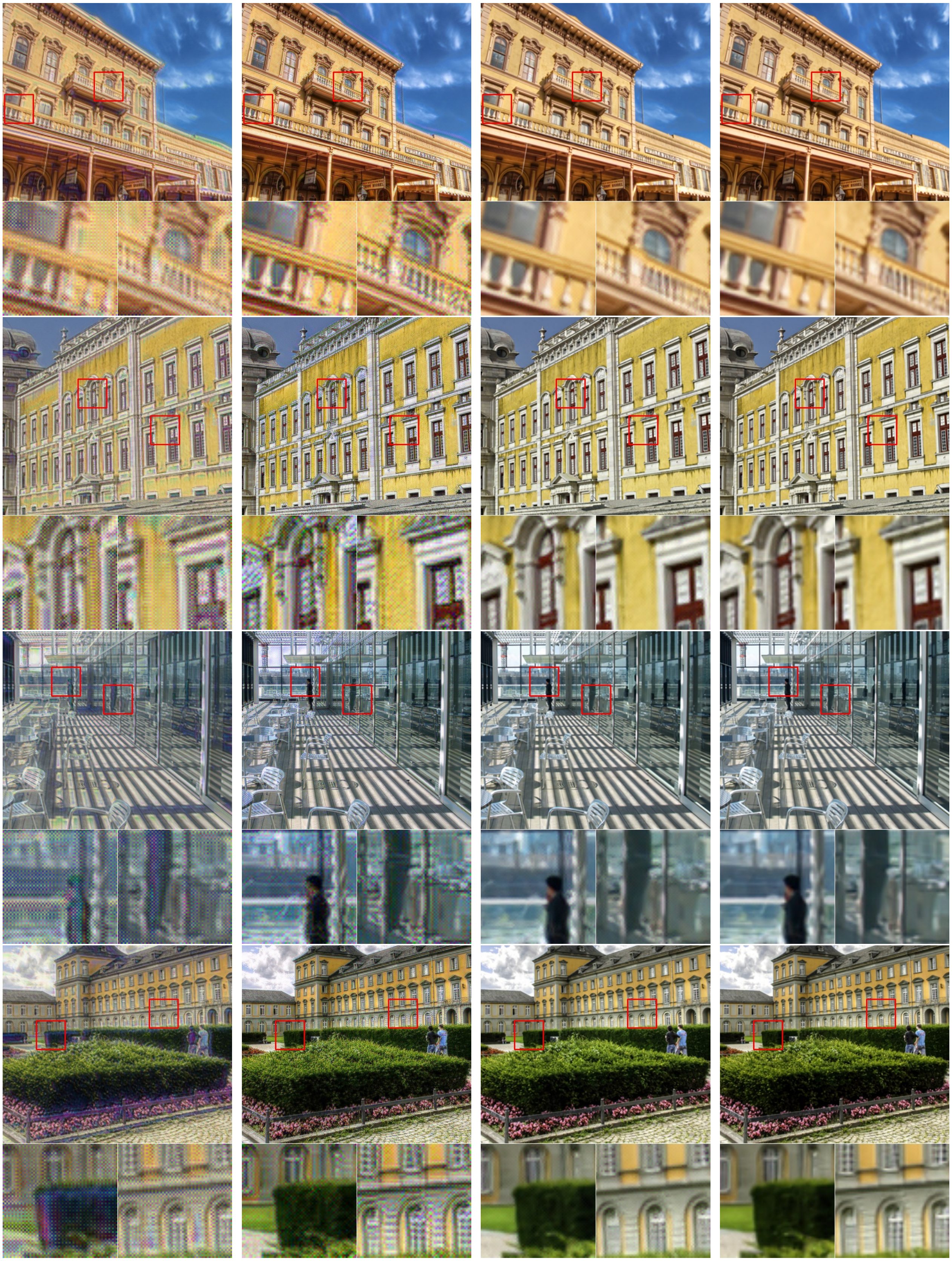
Loss	PSNR↑	LPIPS↓	DISTS↓	SSIM↑	FID↓
L1	21.342	0.206	0.200	0.767	144.866
LPIPS(alex) (0.1)	20.931	0.175	0.174	0.766	87.699
LPIPS(alex) (0.01)	20.824	0.190	0.187	0.763	111.283
CTX (0.1)	21.763	0.134	0.148	0.787	38.779
CTX (0.01)	20.958	0.167	0.174	0.768	88.338
PDL (0.1)	20.797	0.136	0.148	0.747	78.127
PDL (0.01)	20.350	0.150	0.163	0.729	89.211
FDL (0.1)	**20.445**	**0.151**	**0.160**	**0.731**	**91.006**
FDL (0.01)	**21.165**	**0.133**	**0.140**	**0.789**	**30.405**

Table 3. Quantitative comparison of image enhancement using the SwinIR model [16] trained with DPED dataset [13].

mance across all metrics, with only a slight advantage in the PSNR metric. However, when combined with FDL, it exhibits the best performance in terms of LPIPS, DISTS, SSIM, and FID metrics. This finding clearly demonstrates the misalignment robust property of FDL and its ability to ensure the perceptual quality of the predicted results. Furthermore, combining L1 with CTX, which also possesses some misalignment robustness, as a loss function yields the best performance in terms of PSNR. This can be attributed to the fact that CTX calculates the loss function in an element-wise manner on image features, thereby focusing more on local information and emphasizing the preservation of details in the predicted results.

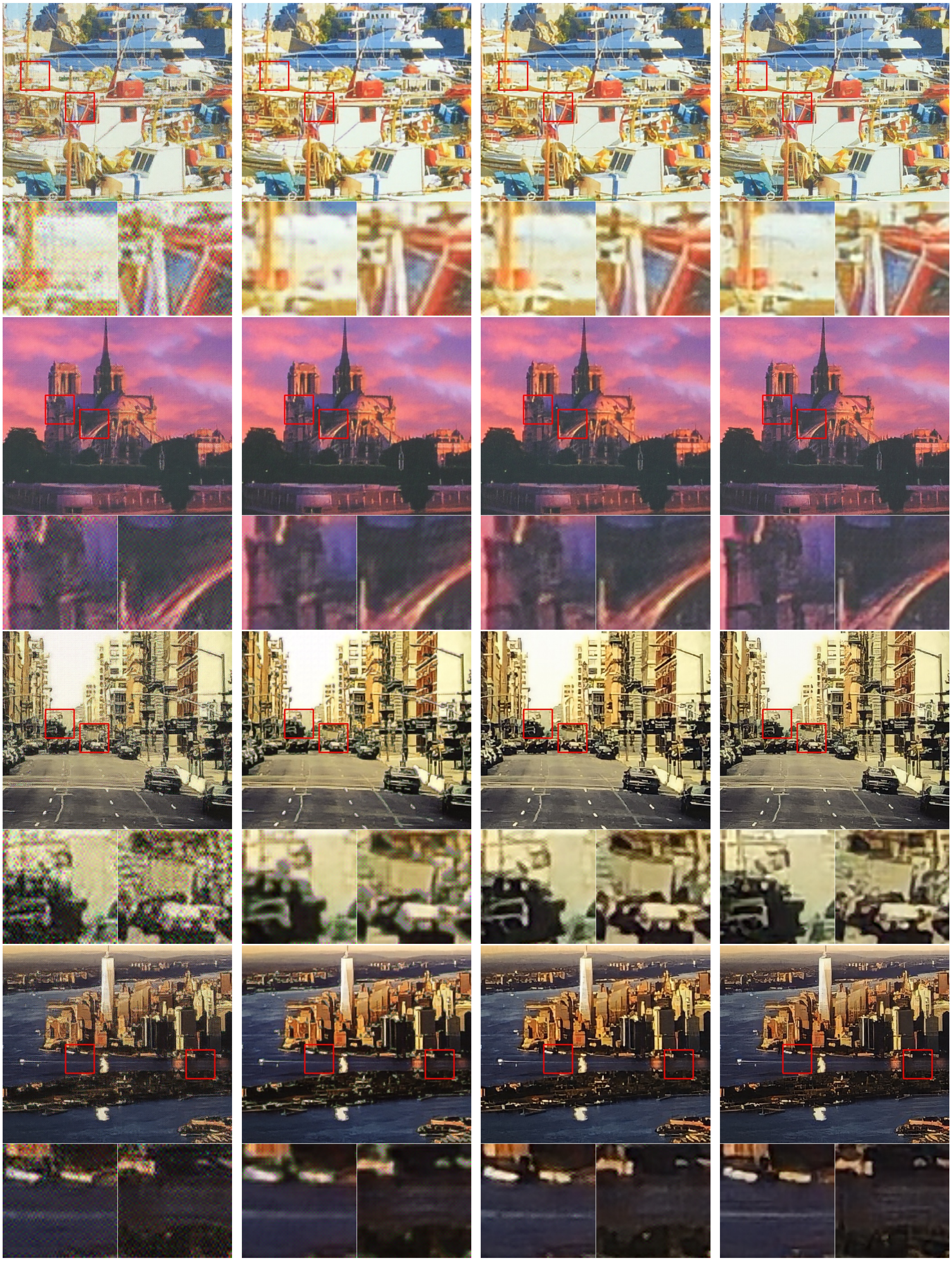
4. Additional Visual Comparisons

This section presents additional qualitative results for various tasks. Figure 6 shows the qualitative results in image enhancement on the DPED dataset. Figure 5 shows the results in SISR on the merged real-world dataset. Additionally, the results of SISR on the synthesized DIV2K with strong misalignments are shown in Figure 4. Finally, Figure 7 shows the result of style transfer compared with Gatys et al [10] and CTX.



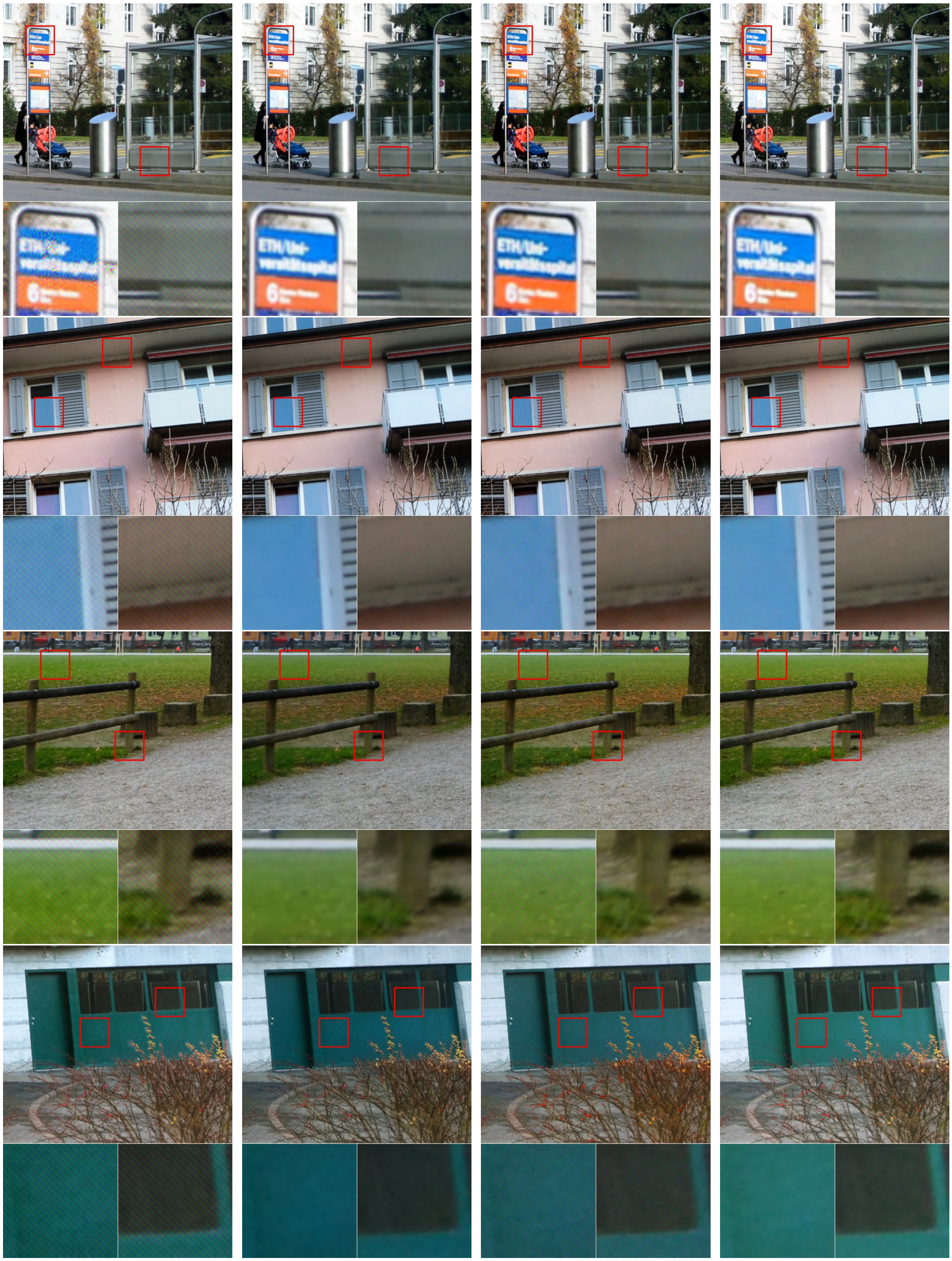
(a) LPIPS (b) PDL (c) CTX (d) FDL (Ours)

Figure 4. Qualitative comparison of SISR using the NLSN model [19] trained with synthetic shifted DIV2K dataset, compared with LPIPS [22], and PDL [7], and CTX [18]. The red area is cropped from different results and enlarged for visual convenient. Zoom in to observe details.



(a) LPIPS (b) PDL (c) CTX (d) FDL (Ours)

Figure 5. Qualitative comparison of real-world SISR using the NAFNet [6] trained with the merged real-world dataset [4, 5].



(a) LPIPS (b) PDL (c) CTX (d) FDL (Ours)

Figure 6. Qualitative results of image enhancement using NAFNet [6] trained with the DPED dataset [13].

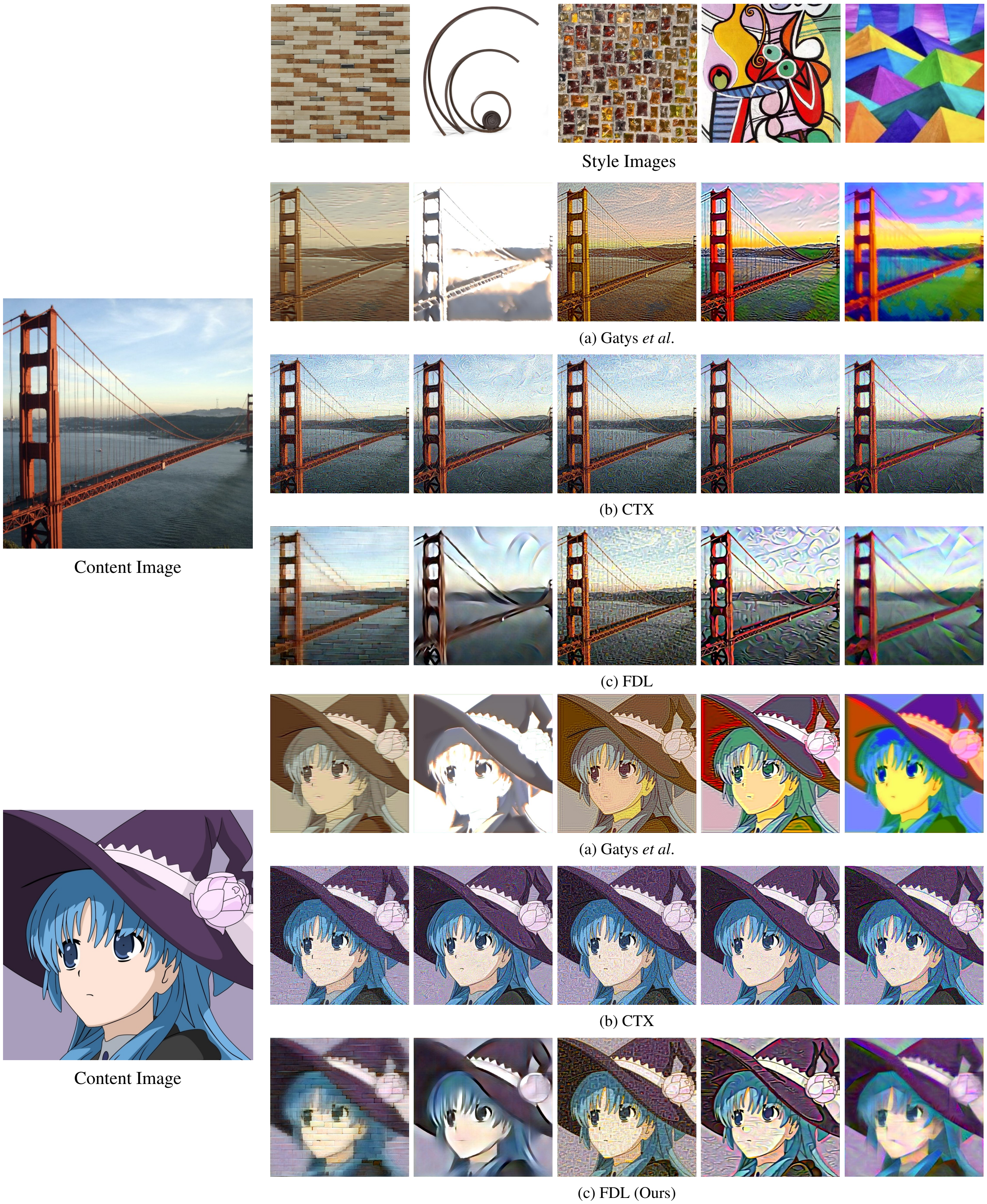


Figure 7. Qualitative results of style transfer compared with Gatys *et al*. and CTX.

References

[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[2] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018.

[3] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2019.

[4] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2019.

[5] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1652–1660, 2019.

[6] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision*, pages 17–33. Springer, 2022.

[7] Mauricio Delbracio, Hossein Talebei, and Pevman Milanfar. Projected distribution loss for image enhancement. In *2021 IEEE International Conference on Computational Photography*, pages 1–12, 2021.

[8] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020.

[9] Ariel Elnekave and Yair Weiss. Generating natural images with direct patch distributions matching. In *European Conference on Computer Vision*, pages 544–560. Springer, 2022.

[10] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[11] Lei Ge and Lei Dou. G-loss: A loss function with gradient information for super-resolution. *Optik*, 280:170750, 2023.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

[13] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. DSLR-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3277–3285, 2017.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[15] 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 Conference on Computer Vision and Pattern Recognition*, pages 1833–1844, 2021.

[16] 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 Conference on Computer Vision and Pattern Recognition*, pages 1833–1844, 2021.

[17] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[18] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *European Conference on Computer Vision*, pages 768–783. Springer, 2018.

[19] 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.

[20] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*. Computational and Biological Learning Society, 2015.

[21] Andrea Vedaldi and Brian Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the ACM International Conference on Multimedia*, pages 1469–1472, 2010.

[22] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.