Misalignment-Robust Frequency Distribution Loss for Image Transformation

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

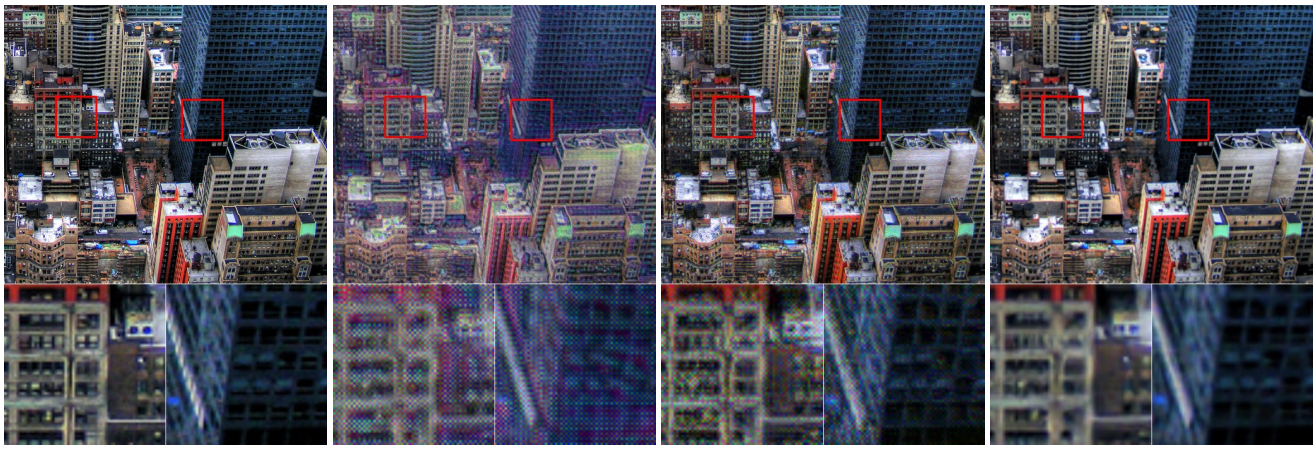


Figure 1. Qualitative results trained on our synthetic DIV2K dataset with strong misalignments. Compared with LPIPS [41], and PDL [8], the proposed method FDL yields clearer results with fewer artifacts. Zoom in to observe details.

Abstract

This paper aims to address a common challenge in deep learning-based image transformation methods, such as image enhancement and super-resolution, which heavily rely on precisely aligned paired datasets with pixel-level alignments. However, creating precisely aligned paired images presents significant challenges and hinders the advancement of methods trained on such data. To overcome this challenge, this paper introduces a novel and simple Frequency Distribution Loss (FDL) for computing distribution distance within the frequency domain. Specifically, we transform image features into the frequency domain using Discrete Fourier Transformation (DFT). Subsequently, frequency components (amplitude and phase) are processed separately to form the FDL loss function. Our method is empirically proven effective as a training constraint due to the thoughtful utilization of global information in the frequency domain. Extensive experimental evaluations, focusing on image enhancement and super-resolution tasks, demonstrate that FDL outperforms existing misalignment-robust loss functions. Furthermore, we explore the potential of our FDL for image style transfer that relies solely on completely misaligned data. Our code is available at: https://github.com/eezkni/FDL

1. Introduction

Image transformation refers to the process of changing the visual appearance or characteristics of images to achieve specific goals or effects. Various studies [10, 15, 29, 31, 43] have showcased impressive results by integrating deep neural networks into image transformation tasks. For example, single image super-resolution (SISR) aims to enhance the spatial properties of images, while image enhancement strives to improve the quality, visibility, interpretability, etc. However, a key limitation of these methods is the implicit assumption of pixel-aligned training data, thereby restricting their scope of applicability. This assumption is problematic as not all image transformation tasks can access perfectly aligned training data, particularly those involving natural distortions. Besides, one prominent example is style transfer, a task that involves optimizing the style distance between images that lack content-related associations. The misalignment of content in training data significantly challenges the effectiveness of these methods.

To mitigate this challenge, a considerable body of re-

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

search has emerged focusing on the development of loss functions to improve the performance of image transformation models, including *element-wise* loss [23, 28, 41] and *distribution-based* loss [8, 11, 18]. Element-wise loss functions are often ill-suited for geometric misaligned training data since even imperceptible misalignment can trigger significant responses to these losses. Mechrez *et al*. [28] tackled the problem of misaligned training data through a patch-matching strategy, resulting in positive outcomes, particularly in misaligned tasks such as semantic style transfer. However, it often introduces artifacts under certain conditions since the spatial structure is ignored [8, 42]. Distribution-based loss functions show promise in mitigating the interference of misaligned data, achieving better perceptual quality, and producing more realistic predictions in misaligned scenarios [8]. However, these distribution-based loss functions often ignore spatial location, leading to potential structural error in predicted results.

In this paper, we present a comprehensive analysis of the distribution distance and propose solutions to address its limitations in accurately capturing the structural integrity of images (as shown in Section. 3.2). Previous work has demonstrated that the frequency domain contains more global information [4, 20, 38, 45]. Our analysis shows that computing distribution distances in the frequency domain, as opposed to the spatial domain, can effectively leverage global information. Consequently, when used as a training constraint, this approach can reduce structural errors in the predicted results. Furthermore, frequency components of images possess various physical meanings [13, 33]. In this work, we observe that the frequency components of image features encompass multiple characteristics within the image. Therefore, integrating information from various frequency components in the loss function can better ensure the overall quality of the predicted images.

As a result, we propose a novel Frequency Distribution Loss (FDL) for image transformation models trained with misaligned data, opening up new avenues for addressing the broad issue of misalignment in image transformation tasks. Specifically, we employ a pre-trained feature extractor to transform images (*i.e.,* predicted image and target image) into the feature space. Subsequently, two frequency components (amplitude and phase) are obtained individually from the predicted image features and the target image features using the Discrete Fourier Transform (DFT). Finally, we employ Sliced Wasserstein Distance (SWD) to measure the distribution distance between the frequency components of predicted and target image features, respectively. We conduct extensive experiments across various image transformation tasks, including single-image super-resolution, image enhancement, and style transfer, to demonstrate the effectiveness of FDL. FDL consistently achieves state-of-the-art performance in all evaluated scenarios, showcasing remarkable robustness to both models and tasks. As illustrated in Figure 1, FDL adeptly assesses the differences among essential information for SISR, even in the presence of strong geometric misalignment, ensuring the comprehensive quality of the predicted image.

2. Related Work

Element-wise Losses for Image Transformation. These loss functions calculate differences between pixels or features of images using an element-wise approach (*e.g.*, L1 or L2 norm, Cosine distance). In many image transformation tasks, this type of loss proves effective in reducing distortion in the predicted image and ensuring its detail fidelity [23, 41, 44]. However, when dealing with misaligned training data, even imperceptible small geometric variations can result in significant responses in such loss functions. This lack of robustness to misalignment can lead to regression to the mean phenomenon in misaligned situations [8], posing challenges in ensuring the quality of predicted images. Therefore, several efforts have been made to enhance the robustness of feature extractors to geometric misalignment, including techniques such as anti-aliasing and max-pooling [16, 39, 40]. These improvements have shown promise, particularly in image quality assessment tasks. However, these modifications to models can lead to information loss, which poses challenges when employing them as loss functions in image transformation tasks. This limitation hinders the assurance of maintaining the quality of predicted images. Mechrez *et al*. proposed Contextual Loss (CTX) by treating image features as a collection of patches and assessing the similarity between two input images through the calculation of element-wise distances between each feature patch and its nearest neighbor [28]. The CTX loss brings a simple solution to misaligned data, however, since CTX cannot effectively utilize the global structural information of the image, artifacts may still appear in the predicted image [42].

Distribution-based Losses for Image Transformation. These loss functions leverage distribution distances or disparities, such as Wasserstein Distance (WD) [11, 30] or Kullback–Leibler divergence (KLD) [7], to quantify the differences between image datasets or instances. Initially used in image generation tasks, these loss functions have gained widespread application in transformation tasks. Empirical evidence has shown a strong correlation between these metrics and the perceptual quality of images [2]. GAN loss can be applied to completely misaligned data, it often introduces artifacts in predicted images because GAN optimizes the distance between two image set-level distributions [31, 32]. In contrast, Elnekave *et al*. [11] addresses the preservation of quality in the predicted image by matching the patch distribution of images. Similarly, PDL [8] cal-

culates the distance between the distribution of image features. These metrics based on spatial domain distribution distance at the image level show robustness to misalignment and can better ensure the quality of images. However, these distribution-based measures only focus on distribution and ignore spatial location information. Therefore, when using distribution distance as the loss function, it is hard to preserve structural accuracy in the predicted results.

3. Methodology

3.1. Overview

We aim to design a loss function tailored for image transformation tasks, capable of measuring the similarity between misaligned images to ensure the overall quality. In Section. 3.2, we conduct a comprehensive analysis of the merits and challenges associated with distribution-based loss functions for image transformation tasks involving misaligned data. We empirically demonstrate that computing distribution distances in the frequency domain can alleviate the challenge of disregarding positional information when calculating distribution distances in the spatial domain, thereby preserving the structural integrity of the predicted results. In Section. 3.3, we explore the diverse information inherent in frequency components of image features. Specifically, we demonstrate that two frequency components in the image feature space (amplitude and phase) are related to various characteristics of the image. Therefore, integrating information in these frequency components is capable of ensuring the quality of the image in various aspects. The overview of our proposed Frequency Distribution Loss (FDL) is shown in Figure 2.

3.2. Frequency Distribution Distance

Wasserstein Distance (WD) has been widely used to optimize neural networks by quantifying the dissimilarity between probability distribution. It completely ignores spatial position information [24], making it robust to geometric misalignment because it focuses on estimating the differences between the underlying distribution of the signals rather than their spatial alignment. However, the disregard for spatial information may lead to the inability of WD to ensure the structural accuracy of the predicted results. We argue that calculating WD in the spatial domain utilizes only the local information while utilizing global information can help address this issue. Therefore, we anticipate that computing WD in the frequency domain better preserves the structural accuracy of predictions due to the richer global information presented in the frequency domain [4, 22, 45].

To validate this hypothesis, we conduct a straightforward toy experiment. Specifically, we generate a set of training data containing multiple pairs of one-dimensional input signals and their corresponding targets, where each pair of input and target signals has slightly different shapes (as shown in Figure 3). Additionally, we introduce random shifts to the target and input signals to induce misalignment within the training pairs. We train a simple model $\mathrm{M}(\cdot)$ to emulate the mapping from source to target, which can be formulated as:

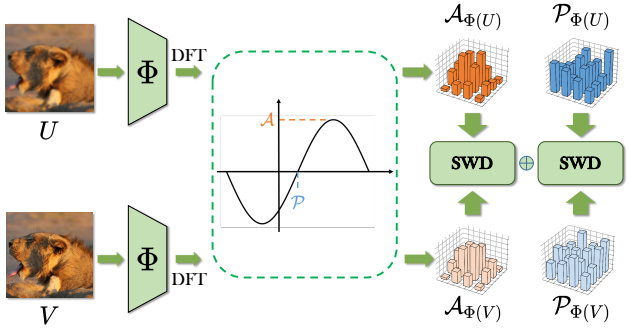


Figure 2. An overview of the proposed Frequency Distribution Loss (FDL). A shared feature extractor network Φ is utilized to project images into perceptual feature space. Subsequently, the amplitude and phase of image features are obtained by Discrete Fourier Transform (DFT). Then, the Sliced Wasserstein Distance (SWD) [11], as an approximation of WD, is performed separately for amplitude and phase, and the results are linearly combined.

$$\mathrm{M}(x) = f(x) + x, \tag{1}$$

where $f(\cdot)$ represents a simple network, calculating the residual between the target and the input single x.

In the one-dimensional scenario, the WD between distribution has a closed-form solution. The loss function based on spatial domain WD can be formulated as:

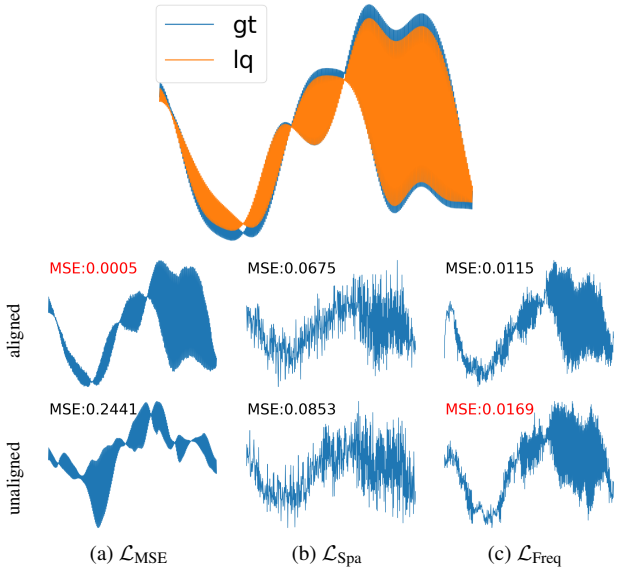
$$\mathcal{L}_{\text{Spa}}(\mathrm{M}(x), y) = \mathrm{WD}\left(\mathrm{M}(x), y\right), \tag{2}$$

where x and y are the input and target signal, respectively, $\mathrm{WD}(\cdot,\cdot)$ represents the one-dimensional Wasserstein Distance between the distribution of the signals. To calculate WD in frequency domain, we initially utilize the Discrete Fourier Transform (DFT) to transform signals into the frequency domain, obtaining frequency components (amplitude and phase), which contain all the frequency domain information. Next, the loss function based on frequency WD can be formulated as:

$$\mathcal{L}_{\text{Freq}}(\mathrm{M}(x), y) = \mathrm{WD}\left(\mathcal{A}_{\mathrm{M}(x)}, \mathcal{A}_y\right) + \mathrm{WD}\left(\mathcal{P}_{\mathrm{M}(x)}, \mathcal{P}_y\right), \tag{3}$$

where $\mathcal{A}_s = |\mathcal{F} \circ s|$ is the amplitude of the spectrum of signal s, and $\mathcal{P}_s = \angle(\mathcal{F} \circ s)$ is the phase, $\mathcal{F}$ denotes DFT. We employ Mean Squared Error ($\mathcal{L}_{\text{MSE}}$), $\mathcal{L}_{\text{Spa}}$ and $\mathcal{L}_{\text{Freq}}$ as loss function to train the mode respectively. And we conduct training with both aligned and misaligned training data.

The comparison results in Figure 3 show that directly using $\mathcal{L}_{\text{MSE}}$ enables the model to effectively learn the map-



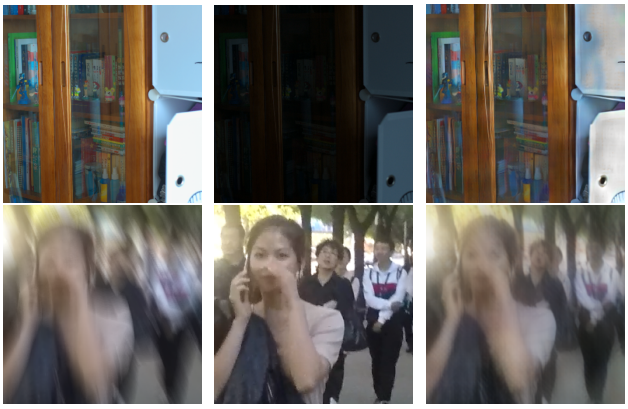
(a) $\mathcal{L}_{\text{MSE}}$ (b) $\mathcal{L}_{\text{Spa}}$ (c) $\mathcal{L}_{\text{Freq}}$

Figure 3. In the one-dimensional scenario, different loss functions are employed to train the same models with aligned and randomly misaligned training data, respectively. lq is the input test signal, and gt is the corresponding ground truth. Each column represents the predicted results of models trained using different loss functions, with the MSE between the predicted result and ground truth.

ping from input to target when there is no misalignment in the training data. In contrast, when there is misalignment in the training data, models trained with $\mathcal{L}_{\text{MSE}}$ exhibit a significant decrease in prediction accuracy compared to perfectly aligned training data. Meanwhile, models trained with $\mathcal{L}_{\text{Spa}}$ and $\mathcal{L}_{\text{Freq}}$ as loss functions show less change in performance. This indicates that both $\mathcal{L}_{\text{Spa}}$ and $\mathcal{L}_{\text{Freq}}$ exhibit shift robustness. However, $\mathcal{L}_{\text{Spa}}$ completely disregards spatial positional information, leading to the model output having a similar distribution to the target but not guaranteeing the structural accuracy of the prediction. Therefore, we turn to measure the WD of the frequency domain for better structural accuracy.

3.3. Feature Frequency Components

In Section. 3.2, we empirically demonstrate the benefit of calculating the distribution distance in frequency domain using amplitude and phase. These frequency components of the image possess specific physical meanings [13, 33], thus we reckon that these frequency components of the image feature are likely to be associated with certain image characteristics. In this section, we further investigate the information associated with amplitude and phase of image features. We provide empirical analysis through a simple experiment. Specifically, we first extract features from two images denoted as Q and D using the encoder $\Phi(\cdot)$ based on VGG. Next, we obtain the amplitude and phase of these



(a) Q (b) D (c) res

Figure 4. Result of frequency components mixing. An encoder (Φ) extracts features from Q and D. We mix the frequency components using the amplitude of $\Phi(Q)$ and the phase of $\Phi(D)$. Then the feature with mixed-frequency component is decoded into the pixel domain.

two features through DFT respectively, and mix the amplitude of $\Phi(Q)$ and the phase of $\Phi(D)$. Subsequently, we project the mixed amplitude and phase back into the feature domain through inverse DFT and adopt a decoder to transform the feature obtained from the mixed frequency components back into pixel space. The pretext process can be expressed as:

$$res = \Phi^{-1}\left(\mathcal{F}^{-1} \circ \left(\mathcal{A}_{\Phi(Q)}, \mathcal{P}_{\Phi(D)}\right)\right), \tag{4}$$

where $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are the encoder and decoder respectively. $\mathcal{F}^{-1}$ denotes the inverse DFT and res is the generated images.

Figure 4 shows the experimental results, where the images are sourced from the LOL [37] and HIDE [34] datasets. By comparing the generated images with image Q and image D, we can discern the information associated with frequency components in the feature domain. Observation shows that the resulting image's texture-related attributes like illumination and color resemble those of image Q, which provides amplitude. Meanwhile, the structural elements such as object shapes and edges exhibit high similarity between the result and image D. We therefore argue that in the feature domain, the amplitude and phase component is associated with information related to various characteristics of the image. Therefore, we believe that incorporating the information in these frequency components in the loss function, allows for a comprehensive consideration of the various characteristics within the image, thereby enhancing the overall quality of the predicted results.

3.4. Overall Loss Function

We summarize the previous analysis and propose the Frequency Distribution Loss (FDL) between the predicted im-

age and the target image, which can be formulated as:

$$\mathcal{L}_{\text{FDL}}\left(U, V\right) = \text{SW}\left(\mathcal{A}_{\Phi(U)}, \mathcal{A}_{\Phi(V)}\right) + \lambda \cdot \text{SW}\left(\mathcal{P}_{\Phi(U)}, \mathcal{P}_{\Phi(V)}\right), \tag{5}$$

where U and V refer to predicted and target images, respectively. $\Phi(\cdot)$ refers to an arbitrary feature extractor. $\text{SW}(\cdot, \cdot)$ represents the Sliced Wasserstein Distance (SWD) between the distribution of two signals. Due to the absence of a closed-form solution for WD in high-dimensional spaces, following the Elnekave *et al*. [11], we employ the SWD as an approximation of WD. As shown in Figure 5, FDL exhibits strong shift invariance, making it suitable for various scenarios with geometric misalignment.

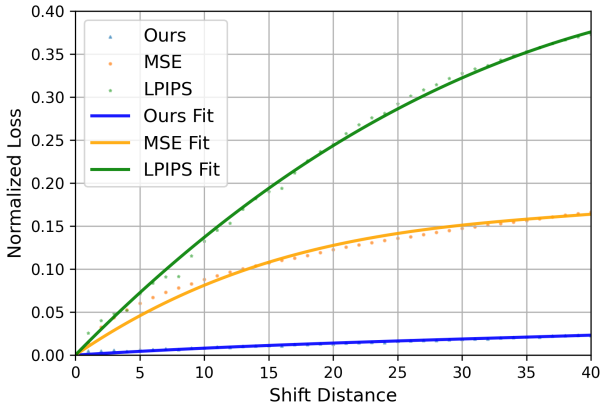


Figure 5. Shift response curves for different loss functions, including FDL, LPIPS, and Mean Square Error(MSE). We randomly shift the reference image for different pixels and calculate the discrepancy between the shifted and reference image using different metrics. The proposed FDL demonstrates strong shift robustness.

4. Experiment

To demonstrate the superiority and generality of our method, we adopt the proposed FDL into various image transformation tasks, including image enhancement, single image super-resolution, and style transfer. For each task, we employ multiple representative baseline models and ensure that each model is trained using only the proposed FDL or the compared loss functions. Note that our focus is exclusively on scenarios where training data is misaligned.

4.1. Experiment Settings

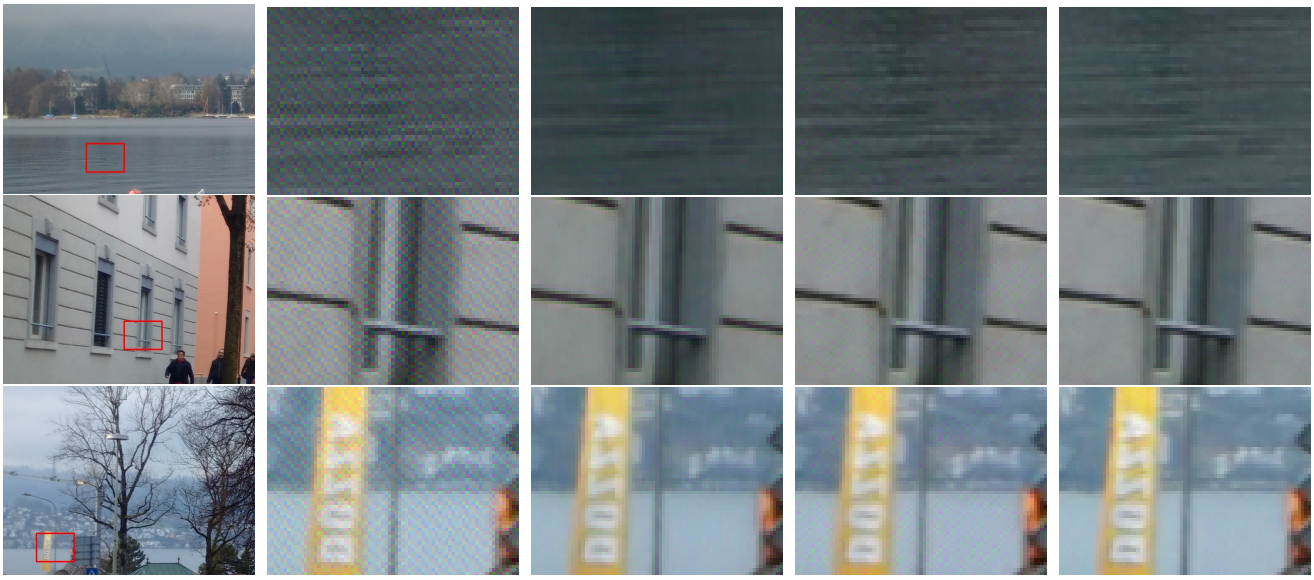
Baseline Models. To comprehensively validate the robustness of our proposed FDL across different architectural models, we select various baseline models for each task: 1) NAFNet [6] and SwinIR [26] for image enhancement; 2) NLSN [29], NAFNet [6] and SwinIR [26] for single image super-resolution (SISR); and 3) Gatys *et al*. [14] for style transfer. The NAFNet, NLSN and Gatys *et al*. are convolutional neural networks (CNN) based models, while SwinIR is a Transformer [27] based model. These models have shown impressive results in corresponding tasks and have been recognized as representative models in recent years.

Baseline Datasets. For image enhancement, we choose the DPED [21] dataset for both training and testing. DPED exhibits significant geometric misalignment between the low-quality images and the high-quality image pairs because these image pairs are captured by different devices with the same scene. Despite employing alignment algorithms and cropping the training images into smaller patches to minimize the impact of misalignment, visually noticeable misalignment still exists in this dataset. Additionally, DPED contains a substantial amount of real-world noise, posing challenges for both the models and the loss functions. For SISR, we select the real-world SISR datasets for training and testing by combing the RealSR [3] and City100 [5] datasets. Furthermore, to examine the capability and generality of FDL in the presence of strong misalignment, a dataset with significant misalignment is synthesized based on the DIV2K dataset. We randomly crop two images with noticeable geometric misalignment from the high-resolution image and downsample one of the cropped images to generate a low-resolution image. This low-resolution image is paired with the other cropped high-resolution image to train SISR models. This is done to simulate irregular displacements that may occur in real-world scenarios.

Baseline Loss Functions. We compare our proposed method with several state-of-the-art loss functions, including CTX [28], PDL [8], and LPIPS [41]. CTX and PDL are loss functions specifically designed for handling misaligned data. LPIPS is a well-known and widely used perceptual loss in various image restoration tasks. All the loss functions follow the official settings for fairness comparison. Specifically, VGG19 [35] is used as the feature extractor for CTX and PDL, while AlexNet [25] is utilized as the feature extractor for LPIPS.

Evaluation Metrics. We select PSNR, SSIM [36], LPIPS [41], DISTS [9], and FID [19] as the evaluation metrics for image enhancement and SISR. SSIM [36] and PSNR can assess the fidelity of image details, while DISTS [9], FID [19], and LPIPS [41] reflect the perceptual quality of predicted images.

Implementation Details. In our work, we employ VGG19 [35] as the feature extractor and compute FDL on the $Relu\_1\_1$, $Relu\_2\_1$, $Relu\_3\_1$, $Relu\_4\_1$, and $Relu\_5\_1$ layers. In different scenarios, we adjust the parameter λ to modulate the weight assigned to different frequency components. Specifically, for super-resolution and style transfer tasks, $\lambda = 1$, while for the image enhancement task, $\lambda = 0.01$.



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

Figure 6. Qualitative results on DPED dataset [21] and NAFNet [6] compared with LPIPS, PDL, CTX. The red area is cropped from different results and enlarged for visual convenient. Zoom in to observe details.

Model	Loss	PSNR↑	LPIPS↓	DISTS↓	SSIM↑	FID↓
NAFNet	CTX	22.256	0.126	0.148	0.778	74.553
	LPIPS(Alex)	20.819	0.291	0.233	0.584	215.350
	PDL	22.665	0.117	0.128	0.776	75.124
	FDL (Ours)	**23.048**	**0.114**	**0.121**	**0.811**	**37.501**
SwinIR	CTX	20.800	0.134	0.152	0.734	64.093
	LPIPS(Alex)	21.613	0.157	0.168	0.759	127.310
	PDL	20.256	0.152	0.167	0.701	107.726
	FDL (Ours)	**21.488**	**0.128**	**0.136**	**0.786**	**29.877**

Table 1. Quantitative comparison of image enhancement on the DPED dataset [21]. The best and second best results are marked in red and blue, respectively.

Model	Loss	PSNR↑	LPIPS↓	DISTS↓	SSIM↑	FID↓
NAFNet	CTX	24.615	**0.245**	0.105	0.833	15.977
	LPIPS(Alex)	16.968	0.441	0.274	0.461	35.440
	PDL	17.737	0.267	0.134	0.595	16.038
	FDL (Ours)	**24.865**	**0.265**	**0.100**	**0.834**	**15.233**
SwinIR	CTX	35.249	0.093	0.101	0.964	66.114
	LPIPS(Alex)	35.198	0.114	0.114	0.958	49.362
	PDL	34.733	0.086	0.094	0.953	35.275
	FDL (Ours)	**35.771**	**0.085**	**0.088**	**0.965**	**23.299**

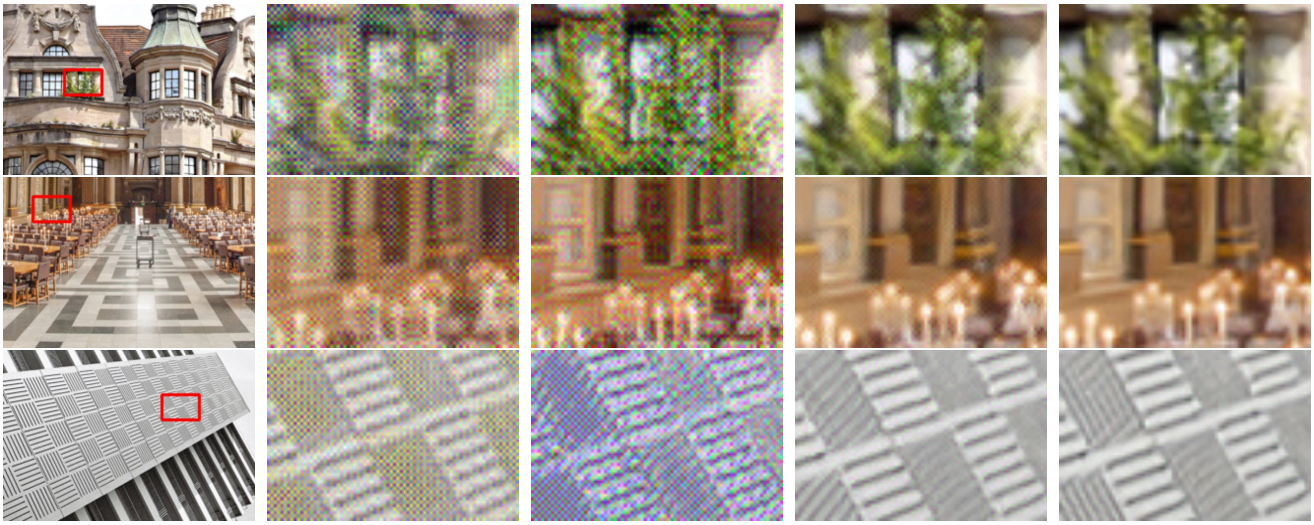
Table 2. Quantitative comparison of SISR on the merged real-world dataset [3, 5].

4.2. Image Enhancement

The quantitative comparison results presented in Table 1 demonstrate the superiority of our proposed FDL over all compared loss functions across various evaluation criteria. This indicates in the presence of significant geometric misalignment in the dataset, our loss function not only preserves fine details in the images but also achieves superior perceptual quality compared to existing misaligned loss functions. Thus, the proposed FDL achieves a better perceptual distortion tradeoff [2]. These advantages can be attributed to the integration of frequency domain information in our loss function. The visual comparison in Figure 6 provides several insightful observations. The element-wise loss functions like CTX and LPIPS struggle to accurately capture differences in structured detail information, resulting in noticeable artifacts. In contrast, distribution based loss functions such as the proposed FDL and PDL can significantly reduce artifacts. The limitations of PDL in accurately measuring more global differences in structural information arise from its calculation of distribution distances in the spatial domain. Our proposed FDL addresses this limitation by calculating distribution distance in the frequency domain, which helps it successfully achieve excellent results in the presence of strong geometric misalignment.

4.3. Super Resolution

We compare our proposed FDL against state-of-the-art loss functions in real-world SISR. Table 2 presents the quantitative results of two representative models (*i.e.*, NAFNet and SwinIR), we can observe that our method outperforms all competing methods almost on all evaluation metrics. On the one hand, our method significantly outperforms the PDL that computes distribution distance in the spatial domain, demonstrating the reasonability of the use of fre-



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

Figure 7. Qualitative comparison results using different loss functions on our synthetic DIV2K dataset [1] with strong misalignment.

quency components in our FDL. On the other hand, our method eliminates the effects of misalignment by utilizing the global structural information in the frequency domain, which still outperforms the CTX. Furthermore, Table 3 reports the comparison results of various loss functions on our synthetic shifted DIV2K dataset with strongly misaligned data. We can clearly observe that our proposed FDL outperforms all competing loss functions by large margins over all four testing set. Compared with CTX, our FDL achieves a substantial improvement in PSNR, increasing from 25.14dB to 26.70dB on the Urban100 test set, while also excelling in perceptual metrics such as LPIPS, DISTS, and FID. This shows that our method achieves a good trade-off between fidelity and quality by measuring distribution distances in the frequency domain. Figure 7 shows the qualitative results of our FDL and the other state-of-the-art methods on the synthetic DIV2K dataset. It is clearly found that the quality of the prediction results of our proposed method is significantly better than that of the comparison method because the results of our method contain less noise and disordered structures.

Test Set	Loss	PSNR↑	LPIPS↓	DISTS↓	SSIM↑	FID↓
Set5	CTX	30.023	0.095	0.092	0.933	5.550
	LPIPS(Alex)	21.754	0.400	0.312	0.438	66.619
	PDL	29.598	0.114	0.095	0.767	7.682
	FDL (Ours)	**32.478**	**0.092**	**0.093**	**0.950**	**3.853**
Set14	CTX	27.836	0.156	0.105	0.938	7.220
	LPIPS(Alex)	21.019	0.401	0.320	0.423	50.127
	PDL	26.832	0.165	0.118	0.702	9.727
	FDL (Ours)	**29.526**	**0.152**	**0.103**	**0.957**	**5.853**
B100	CTX	27.829	0.152	0.114	0.881	16.681
	LPIPS(Alex)	22.041	0.367	0.299	0.390	91.150
	PDL	27.231	0.159	0.123	0.645	17.308
	FDL (Ours)	**28.968**	**0.150**	**0.110**	**0.902**	**20.451**
Urban100	CTX	25.138	0.143	0.104	0.850	8.582
	LPIPS(Alex)	19.987	0.382	0.297	0.369	79.639
	PDL	23.847	0.194	0.136	0.572	22.389
	FDL (Ours)	**26.702**	**0.137**	**0.098**	**0.887**	**7.903**

Table 3. Quantitative comparison of NLSN [29] for SISR on our synthetic shifted DIV2K dataset [1].

4.4. Style Transfer

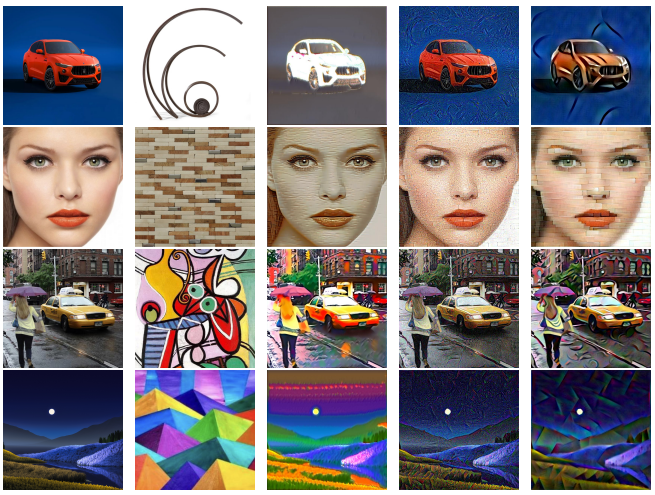
Style transfer aims to synthesize a new image that combines the content of one image with the artistic or stylistic features of another. The primary challenge of style transfer is finding the right balance between preserving the content of the input image and incorporating the stylistic features from the reference style image. Following the pipeline of Gatys *et al*. [14], we optimize the generated image with content loss and style loss. Our proposed FDL is also capable of handling this challenging task, since the use of distribution distance measurement in the frequency domain. Specifically, we define content loss and style loss as follows:

$$\mathcal{L}_{\text{style}}(R, S) = \mathcal{L}_{\text{FDL}}(R, S), \tag{6}$$

$$\mathcal{L}_{\text{content}}(R, T) = \text{SW}\left(\mathcal{P}_{\Phi(R)}, \mathcal{P}_{\Phi(T)}\right), \tag{7}$$

where R is the generated image, S and T refer to style and content image, respectively. We compare FDL with CTX and perceptual losses in Gatys *et al*., and use their respective official settings for fair comparison.

Visual comparison results are presented in Figure 8. It can be observed that losses in Gatys *et al*. [14] only capture the color information from the style image, resulting in poor performance in transferring structured styles. On the other hand, CTX focuses on local texture patterns in the style im-



(a) Content (b) Style (c) Gatys *et al*. (d) CTX (e) FDL (Ours)

Figure 8. Qualitative comparison results compared with Gatys *et al*. and CTX. Our FDL loss function can better retain the structural information in style images.

Loss	PSNR↑	LPIPS↓	DISTS↓	SSIM↑	FID↓
$\mathcal{L}_{\text{spatial}}$	22.916	0.118	0.125	0.798	61.087
$\mathcal{L}_{\text{FDL}}$	23.048	0.114	0.121	0.811	37.501

Table 4. Ablation of calculating in the frequency domain.

age but fails to achieve style transfer at a more global and structural level. In contrast, our method can effectively capture the structural information present in the style image. This demonstrates the effectiveness of utilizing frequency domain global information.

4.5. Ablation Study

We conduct a series of ablation experiments on the NAFNet in the DPED dataset. Firstly, to validate the impact of computing distribution distance in the frequency domain, we calculate the SWD [11] as the loss function in the image feature spatial domain, as follows:

$$\mathcal{L}_{\text{spatial}}(U, V) = \text{SW}\left(\Phi(U), \Phi(V)\right). \tag{8}$$

From Table 4, compared with $\mathcal{L}_{\text{spatial}}$, we can observe that the proposed FDL shows improvements across all metrics. This observation suggests that computing the distribution distance between global information in the frequency domain as a loss function can better ensure the overall quality of the predicted results.

Next, we aim to further investigate the effect of different feature extractors on the final results. To achieve this, we conducted experiments by using ResNet [17] and EffNet [12] as feature extractors in our proposed loss. In particular, we also remove the feature extractor in the proposed loss function, thereby directly performing FDL loss

Loss	Backbone	PSNR↑	LPIPS↓	DISTS↓	SSIM↑	FID↓
$\mathcal{L}_{\text{FDL}}(\lambda = 0.01)$	ResNet	22.415	0.139	0.146	0.763	79.812
$\mathcal{L}_{\text{FDL}}(\lambda = 0.01)$	EffNet	20.389	0.196	0.178	0.578	173.891
$\mathcal{L}_{\text{FDL}}(\lambda = 0.01)$	None	22.581	0.149	0.156	0.789	60.110
$\mathcal{L}_{\text{FDL}}(\lambda = 100)$	VGG	22.134	0.131	0.141	0.787	67.766
$\mathcal{L}_{\text{FDL}}(\lambda = 10)$	VGG	22.815	0.121	0.128	0.803	52.503
$\mathcal{L}_{\text{FDL}}(\lambda = 1)$	VGG	22.810	0.117	0.124	0.806	55.629
$\mathcal{L}_{\text{FDL}}(\lambda = 0.1)$	VGG	22.822	0.118	0.126	0.804	65.711
$\mathcal{L}_{\text{FDL}}(\lambda = 0.01)$	VGG	23.048	0.114	0.121	0.811	37.501

Table 5. Ablation of backbone feature extractor and the weight of SWD between phase component of features.

calculation on pixels (i.e., the "None" in Table 5). The quantitative results of different feature extractors are shown in Table 5, and we can observe that VGG19 [35] yields the best performance among all the results.

Finally, we aim to explore the impact of the weight assigned to the distribution distance between amplitude and phase, by adjusting λ in Equation 5. Table 5 reports the comparison results of different settings of λ, and we can observe that $\lambda = 0.01$ performs best among all settings. This can be attributed to the fact that in the DPED [21] dataset, the main difference between the input and target images lies in the texture, which is highly correlated with the amplitude of the image features. Therefore, assigning a higher weight to the amplitude component in FDL helps the model achieve better performance on the DPED dataset. This observation suggests that in image transformation tasks with different emphasis on different image characteristics, adjusting the value of λ allows the model to allocate different priorities to the characteristics of the predicted results.

5. Conclusion

This paper proposes a robust misalignment loss for image transformation tasks. Our proposed FDL calculates the distribution distance in the frequency domain of image features. Through experiments, we have demonstrated that frequency domain components of image features contain global information closely related to multiple image characteristics. By utilizing the distance between the distribution of these global information as a loss function, we can mitigate the limitation in spatial distribution distances, and ensure the overall quality of the predicted results. In future work, we hope to investigate the frequency components of image features further and improve the performance of FDL by assigning different attention weights to distinct frequency domain regions.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62201387 and 62371343, in part by the Shanghai Pujiang Program under Grant 22PJ1413300, and in part by the Fundamental Research Funds for the Central Universities.

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. 7

[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. 2, 6

[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. 5, 6

[4] Mu Cai, Hong Zhang, Huijuan Huang, Qichuan Geng, Yixuan Li, and Gao Huang. Frequency domain image translation: More photo-realistic, better identity-preserving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13930–13940, 2021. 2, 3

[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. 5, 6

[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. 5, 6

[7] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018. 2

[8] 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. 1, 2, 5

[9] 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. 5

[10] 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. 1

[11] Ariel Elnekave and Yair Weiss. Generating natural images with direct patch distributions matching. In *European Conference on Computer Vision*, pages 544–560. Springer, 2022. 2, 3, 5, 8

[12] Ido Freeman, Lutz Roese-Koerner, and Anton Kummert. Effnet: An efficient structure for convolutional neural networks. In *IEEE International Conference on Image Processing*, pages 6–10. IEEE, 2018. 8

[13] Carl M Gaspar and Guillaume A Rousselet. How do amplitude spectra influence rapid animal detection? *Vision Research*, 49(24):3001–3012, 2009. 2, 4

[14] 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. 5, 7

[15] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *In ACM Transactions on Graphics*, 36(4):1–12, 2017. 1

[16] Abhijay Ghildyal and Feng Liu. Shift-tolerant perceptual similarity metric. In *European Conference on Computer Vision*, pages 91–107. Springer, 2022. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 8

[18] Eric Heitz, Kenneth Vanhoey, Thomas Chambon, and Laurent Belcour. A sliced Wasserstein loss for neural texture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9412–9420, 2021. 2

[19] 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. 5

[20] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. Deep fourier-based exposure correction network with spatial-frequency interaction. In *European Conference on Computer Vision*, pages 163–180. Springer, 2022. 2

[21] 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. 5, 6, 8

[22] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13919–13929, 2021. 3

[23] 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. 2

[24] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017. 3

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012. 5

[26] 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. 5

[27] 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. 5

[28] 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. 2, 5

[29] 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. 1, 5, 7

[30] Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2020. 2

[31] Zhangkai Ni, Wenhan Yang, Shiqi Wang, Lin Ma, and Sam Kwong. Towards unsupervised deep image enhancement with generative adversarial network. *IEEE Transactions on Image Processing*, 29:9140–9151, 2020. 1, 2

[32] Zhangkai Ni, Wenhan Yang, Hanli Wang, Shiqi Wang, Lin Ma, and Sam Kwong. Cycle-interactive generative adversarial network for robust unsupervised low-light enhancement. In *Proceedings of the ACM International Conference on Multimedia*, pages 1484–1492, 2022. 2

[33] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981. 2, 4

[34] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5572–5581, 2019. 4

[35] 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. 5, 8

[36] 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. 5

[37] Chen Wei, Wenjing Wang, yang Wenhan, and Liu Jiaying. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018. 4

[38] Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2

[39] Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning*, pages 7324–7334, 2019. 2

[40] 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. 2

[41] 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. 1, 2, 5

[42] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2019. 2

[43] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision*, pages 286–301, 2018. 1

[44] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2016. 2

[45] Man Zhou, Jie Huang, Keyu Yan, Hu Yu, Xueyang Fu, Aiping Liu, Xian Wei, and Feng Zhao. Spatial-frequency domain information integration for pan-sharpening. In *European Conference on Computer Vision*, pages 274–291. Springer, 2022. 2, 3