Towards Progressive Multi-Frequency Representation for Image Warping

Jun Xiao*,[1] Zihang Lyu*,[1] Cong Zhang[1] Yakun Ju[2] Changjian Shui[3] Kin-Man Lam†,[1]
[1]The Hong Kong Polytechnic University [2]Nanyang Technological University [3]Vector Institute
{jun.xiao, zihang.lyu, cong-clarence.zhang}@connect.polyu.hk, kin.man.lam@polyu.edu.hk
yakun.ju@ntu.edu.sg, changjian.shui@vectorinstitute.ai

Abstract

Image warping, a classic task in computer vision, aims to use geometric transformations to change the appearance of images. Recent methods learn the resampling kernels for warping through neural networks to estimate missing values in irregular grids, which, however, fail to capture local variations in deformed content and produce images with distortion and less high-frequency details. To address this issue, this paper proposes an effective method, namely MFR, to learn Multi-Frequency Representations from input images for image warping. Specifically, we propose a progressive filtering network to learn image representations from different frequency subbands and generate deformable images in a coarse-to-fine manner. Furthermore, we employ learnable Gabor wavelet filters to improve the model's capability to learn local spatial-frequency representations. Comprehensive experiments, including homography transformation, equirectangular to perspective projection, and asymmetric image super-resolution, demonstrate that the proposed MFR significantly outperforms state-of-the-art image warping methods. Our method also showcases superior generalization to out-of-distribution domains, where the generated images are equipped with rich details and less distortion, thereby high visual quality. The source code is available at https://github.com/junxiao01/MFR.

1. Introduction

Image warping aims to change the appearance of images by performing geometric transformations, which involves changing the positions of image pixels to new positions in a predefined coordinate systems. As a basic technique in computer vision, image warping has become an indispensable component in numerous vision tasks, such as facial manipulation [11, 46], image registration [12, 20, 31], image synthesis [1, 52], etc., substantially affecting their overall performance.

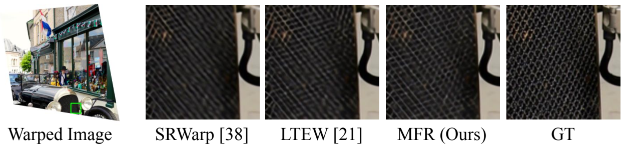

Figure 1. Illustration of local image patches generated by SRWarp [41], LTEW [22], and our proposed MFR.

Traditional image warping methods usually apply an inverse transformation function to deform images and depend on interpolation techniques, such as bicubic interpolation, to estimate the missing values in irregular grids. However, previous studies [9, 30, 36, 41] have revealed that these interpolation-based methods often introduce undesirable jagging and blurry artifacts, leading to image quality degradation. Recent works [22, 41] have formulated image warping as a generalized image super-resolution (SR) problem with varying scaling factors in the spatial domain. This is equivalent to stretching local image regions with different scaling factors in different directions. To generate deformable image content, Son *et al*. [41] employed a pretrained SR model for feature extraction and introduced adaptive warping layers to model the resampling kernel. On the other hand, Lee *et al*. [22] treated image representation in a continuous space and utilized the coordinate-based MLP model to synthesize content in irregular grids. Nevertheless, we observe that these methods encounter challenges in captured local variations in the deformable images, and their generated images are often distorted and lack high-frequency details, as demonstrated in Fig. 1. Furthermore, these methods exhibit poor generalization performance in out-of-distribution data, i.e., scaling factors and geometric transformation not included in the training dataset, significantly limiting their real-world applications. Consequently, there is still substantial potential for improvement in image warping.

*These authors contributed equally to this work
†The corresponding author

In this paper, we propose an effective approach to learn multi-frequency representations from an input image for image warping, namely MFR. Concretely, a progressive filtering network is devised to learn the image representations from different frequency subbands in the feature space. Moreover, the proposed model starts from the input coarse-scale images and gradually produces finer-scale details through the learned frequency representations, resulting in a coarse-to-fine generation process. As image warping intrinsically involves significant local deformation, we incorporate learnable Gabor wavelet filters to improve our model's learning capability of spatial-frequency representations, particularly beneficial for handling high-frequency information in local regions. Extensive experiments have shown that MFR can remarkably outperform state-of-the-art image warping methods in various vision tasks, including homography transformation, equirectangular projection (ERP) to perspective projection, and asymmetric image super-resolution. Notably, our model demonstrates superior generalization capability to out-of-distribution data, yielding images with rich details and high visual quality.

2. Related Works

2.1. Image Warping

Image warping serves as a foundational technique in computer vision and has been widely used in various vision tasks such as image registration [12, 20, 31], image generation [1, 52], and image editing [11, 46]. This technique employs geometric transformations to map pixel positions of images to new locations in a distinct coordinate system. Traditional image warping methods [3, 5, 6, 13, 21] rely on interpolation approaches to compute missing values in irregular grids, but tend to introduce jagged and blurred artifacts, resulting in suboptimal performance. A recent approach by Son *et al*. [41] considers image warping as a generalized image SR problem with varying scaling factors in local regions. They introduced an adaptive warping layer to generate transformed images by leveraging the features extracted from pretrained SR models. Despite its merits, this method exhibits poor generalization in generating large-scale images. To address this issue, Lee *et al*. [22] provided an alternative method that treats image warping as image representation in continuous spaces. They incorporated Fourier features of the coordinate information with the SR features and employed a coordinate-based MLP model to synthesize the warped images. However, this MLP model has limited model capability in feature representation and fails to capture local variations in deformable content, resulting in distortion in the generated images. In contrast, our method focuses on learning local high-frequency representations for effectively generating high-frequency details in deformable regions.

2.2. Learning in the Frequency Domain

Image representations in the frequency domain typically contain distinct patterns and have demonstrated their effectiveness in numerous vision tasks, including image classification [33–35, 43, 49], domain generalization [8, 16, 29, 39, 50], and image generation [4, 10, 18, 19, 25, 28, 37, 42, 47, 48]. Notably, Huang *et al*. [16] introduced a randomization technique in the frequency domain to learn domain-invariant features for domain generalization. Similarly, Yang *et al*. [50] swapped the low-frequency spectrum between the source and target domains to achieve domain alignment, enhancing the model's generalization capabilities. For the generation of high-quality images, Tancik *et al*. [42] introduced Fourier features to enhance the representation ability of the implicit neural representation methods. Sitzmann *et al*. [40] proposed the Sinusoidal Representation Network (SIREN) in which the ReLU function is replaced with a sinuous activation function, demonstrating the potential of frequency representation learning. Inspired by the success of frequency learning in various applications, we propose an effective method to learn frequency components from different frequency subbands in the latent space specifically for image warping. Additionally, to capture local variations in deformable image content, we propose using learnable Gabor wavelet filters to extract spatial-frequency representations from local regions. This approach can significantly enhances the performance by generating clear high-frequency details in the locally deformable areas.

3. Methodology

Given an input RGB image $I_{\text{in}} \in \mathbb{R}^{W\times H\times 3}$, our objective is to synthesize high-quality deformable image content by learning multi-frequency representations from the input, where W and H represent the width and height of the images, respectively. The overall pipeline of our proposed MFR is shown in Fig. 2. It consists of two stages: the feature encoding stage and the frequency learning stage, which are elaborated in the following sections.

3.1. Feature Encoding

In the feature encoding stage, we employ a pretrained SR model for projecting the input images to the latent feature space. Subsequently, we leverage the local texture estimator (LTE) [22] to integrate both coordinate information (i.e., relative position) and geometric information (i.e., curvature) into the output of the pretrained SR model. The overall process of feature encoding is denoted as E_ϕ. This incorporation results in latent representations of the input image, denoted by $X \in \mathbb{R}^{W\times H\times d}$, where d is the dimensionality of the latent space. However, these representations alone are insufficient to capture local variations for image

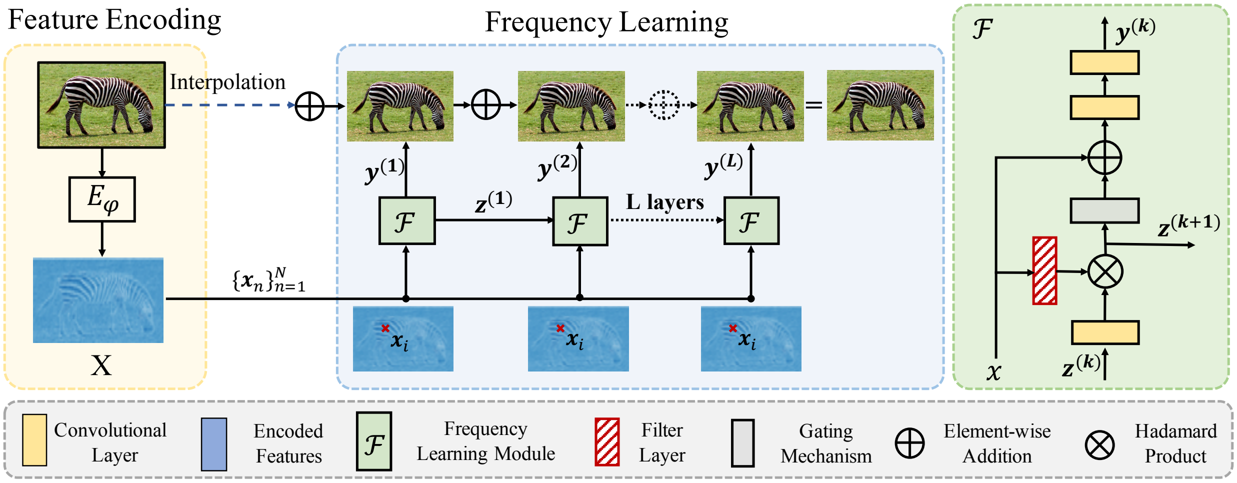

Figure 2. The overall pipeline of the proposed MFR, which consists of two stages: feature encoding and frequency learning.

synthesis.

3.2. Frequency Learning

Given the latent features X obtained in the feature encoding stage, we vectorized them into a set of latent vectors, expressed as $\{\boldsymbol{x}_n\}_{n=1}^N$, where $N = H \times W$ is the total number of latent vectors. Our proposed MFR contains a frequency learning network which comprises L frequency learning modules stacked sequentially. Each module learns the frequency representations from the latent vectors and generates the corresponding deformed image content. Specifically, we assume that $\boldsymbol{x}_i$ is the latent vector at the i-th pixel position. MFR extracts frequency representations $\boldsymbol{z}$ of the input latent vectors through a sequence of frequency learning modules $\mathcal{F}_\ell$, as follows:

$$\boldsymbol{z}_i^{(0)} = \mathcal{G}_{\theta_0}(\boldsymbol{x}_i) + \boldsymbol{x}_i, \tag{1}$$

$$\boldsymbol{z}_i^{(\ell+1)} = \mathcal{G}_{\theta_\ell}(\boldsymbol{x}_i) \otimes \left(W^{(\ell)}\boldsymbol{z}_i^{(\ell)} + \boldsymbol{b}^{(\ell)}\right), \tag{2}$$

for $\ell = 0, \cdots, L-1$. $\boldsymbol{z}_i^{(\ell)} \in \mathbb{R}^d$ represents the frequency representation generated from the ℓ-th module and the input latent vector $\boldsymbol{x}_i$. $\mathcal{G}_{\theta_\ell} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ denotes the filter layer in the ℓ-th module. $W^{(\ell)} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{b}^{(\ell)} \in \mathbb{R}^d$ represent the weights and bias terms of the ℓ-th layer, respectively. $\otimes$ denotes the element-wise multiplication. The proposed MFR extracts frequency representations from the input using the filtering layer and embeds them into the frequency representation obtained from the previous module, to form new frequency representations, resulting in a progressive learning procedure.

When the filter layer adopts a sinusoidal function like SIREN [40], the associated frequency representations can be express as $\boldsymbol{z}_i^{(\ell)} = \sin(\omega_\ell \boldsymbol{x}_i + \phi_\ell)$, where ω_ℓ and ϕ_ℓ are the filer frequency and phase in the ℓ-th module. A larger value of ω_ℓ means that the corresponding module tends to learn a higher frequency response from the input vector. Yüce *et al.* [51] revealed that the network output can be approximated by a polynomial function using the Taylor expansion. As a result, the L-th frequency representation $\boldsymbol{z}_i^{(L)}$ can be equivalently expressed as a linear combination of the sinusoidal functions, as follows:

$$\boldsymbol{z}_i^{(L)} = \sum_{j=0}^{N_{\text{sine}}} \hat{\alpha}_j \sin(\hat{\omega}_j \boldsymbol{x}_i + \hat{\phi}_j), \tag{3}$$

where the parameters $\hat{\alpha}_j$, $\hat{\omega}_j$, and $\hat{\phi}_j$ are dependent on the parameters of the network. N_{sine} denotes the total number of summation terms and is defined as follows:

$$N_{\text{sine}} = \sum_{i=0}^{L-1} 2^i (d)^{i+1}, \tag{4}$$

Obviously, N_{sine} grows exponentially with the number of frequency learning modules used in our MFR. This means that we can use a small number of frequency learning modules to obtain various frequency components of the input.

In each frequency learning module, we compute the output $\boldsymbol{y}_i^{(\ell)}$ based on the associated frequency representation $\boldsymbol{z}_i^{(\ell)}$ and the input $\boldsymbol{x}_i$ as follows:

$$\boldsymbol{y}_i^{(\ell)} = f_{\phi_\ell}(g_\ell(\boldsymbol{z}_i^{(\ell)}) + \boldsymbol{x}_i), \tag{5}$$

where g_ℓ is the gating mechanism in the ℓ-th module, which is a 1-D attention mechanism [14] and adaptively controls the information flow from the frequency representation $\boldsymbol{z}_i^{(\ell)}$. f_{ϕ_ℓ} is the decoding function parametrized by ϕ_ℓ in the ℓ-th module. At the output part of each module, we employ a short connection to fuse the spatial representations $\boldsymbol{x}_i$ with the learned frequency representation $\boldsymbol{z}_i^{(\ell)}$.

After obtaining the output of each module, our model computes the final output image $\boldsymbol{y}$ as follows:

$$\boldsymbol{y} = \boldsymbol{y}_{\text{bic}} + \sum_{\ell=1}^{L} \boldsymbol{y}^{(\ell)}, \tag{6}$$

where $\boldsymbol{y}_{\text{bic}}$ represents the coarse-scale image content obtained by performing bicubic interpolation to the input image. Therefore, our model progressively generates fine-scale image content and enhances the image quality.

3.3. Gabor Wavelet Filter Layer

The primary challenge in generating high-quality deformable images is to effectively capture local intensity variations within deformable content. To facilitate this, the filtering layers used in our model play a crucial role. Instead of using conventional filters like SIREN [40] and Gaussian filters, we incorporate 1-D Gabor wavelet filters into the frequency learning modules which is defined as follows:

$$\mathcal{F}_{\text{Gabor}}(\boldsymbol{x}) = e^{-\frac{(\boldsymbol{x}-\boldsymbol{x}_0)^2}{\alpha^2}} e^{-i\omega(\boldsymbol{x}-\boldsymbol{x}_0)}. \tag{7}$$

The Gabor wavelet filter is a Gaussian filter modulated by a complex exponential term, which has demonstrated remarkable capability to capture local variations in both the spatial and frequency domains in image processing [17, 23, 24, 26, 32, 38]. In Eq. (7), $\boldsymbol{x}_0$ is the predefined center point. As the input $\boldsymbol{x}$ deviates from this center point, the output response undergoes exponential decay. Similarly, we can generate different output responses by controlling the rate of exponential drop-off α and the rate of the modulation ω. Our MFR can learn different frequency representations by learning different values of these two parameters through backpropagation. It is difficult for conventional filters, such as SIREN [40] and Gaussian filters, to capture local variations from both the spatial and frequency domains simultaneously. In contrast, Gabor wavelet filters are defined over compact supports in both the spatial and frequency domains as shown in Fig. 3. As a result, it can effectively capture intensity variations in local regions from both the spatial and frequency domains for image warping.

4. Experiments and Analysis

4.1. Experiment Settings

Dataset information and implementation settings. The proposed MFR was trained using DIV2K dataset [2], which is a widely-used dataset in numerous low-level vision tasks and provides 800 high-resolution images. In the training stage, we randomly crop local image patches of size 48×48 from the input images as the training samples, and set the batch size to 16. We used the ℓ_2 loss as the loss function.

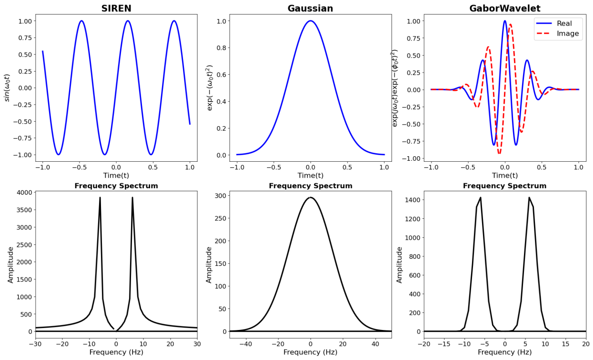

Figure 3. Illustration of three 1-D filters: SIREN [40], Gaussian, and GaborWavelet filters.

The Adam optimizer was utilized to update the model parameters with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the total number of training epochs was 600. We initially set the learning rate to 2×10^{-4} and adaptively decay it by utilizing the cosine annealing strategy.

Evaluation settings. To assess the effectiveness of our model in image warping, we evaluate the model on three vision tasks, including homography transformation, ERP to perspective projection, and asymmetric image super-resolution. The masked peak signal-to-noise ratio (mPSNR) [41] is adopted as the performance metric in the homography transformation task. In asymmetric image super-resolution, the PSNR is used to evaluate the reconstruction quality with different scaling factors.

4.2. Experiments on Homography Transformation

We compare our MFR with SRWarp [41] and LTEW [22] on benchmark datasets provided in [22] in both the in-scale setting and the out-of-scale settings. In the in-scale setting, the scaling factors are involved in the training dataset. However, in the out-of-scale setting, the scaling factors are not included in the training dataset. To ensure a fair comparison, we adopt RRDB [45] as the SR backbone for feature extraction in these three models. For completeness, we include the results from bicubic interpolation and the original RRDB model. For the RRDB model, we first use the model to generate the images, which are then resampled by bicubic interpolation. For SRWarp and LTEW, we directly use their public source codes for implementation.

Table 1 shows the average mPSNR scores of different image warping methods for homograph transformation on the benchmark datasets with two evaluation settings. It is obvious that our MFR achieves better performance than the compared models in the in-scale setting, especially in the Urban100W dataset, and the performance of MFR is 0.18dB higher than the LTEW method. In the out-of-scale setting, the proposed MFR significantly outperforms the

Table 1. The average mPSNR of different image-warping methods for homograph transformation on benchmark datasets with the in-scale (is) and out-of-scale (os) settings. The best results are highlighted in bold.

Methods	DIV2KW		Set5W		Set14W		B100W		Urban100W	
	is	os	is	os	is	os	is	os	is	os
Bicubic	27.85	25.03	35.00	28.75	28.79	24.57	28.67	25.02	24.84	21.89
RRDB [45]	30.76	26.84	37.40	30.34	31.56	25.95	30.29	26.32	28.83	23.94
SRWarp-RRDB [41]	31.04	26.75	37.93	29.90	32.11	25.35	30.48	26.10	29.45	24.04
LTEW-RRDB [22]	31.10	26.92	38.20	31.07	32.15	26.02	30.56	26.41	29.50	24.25
MFR-RRDB (Ours)	**31.18**	**27.12**	**38.23**	**31.19**	**32.26**	**26.26**	**30.62**	**26.53**	**29.68**	**24.51**

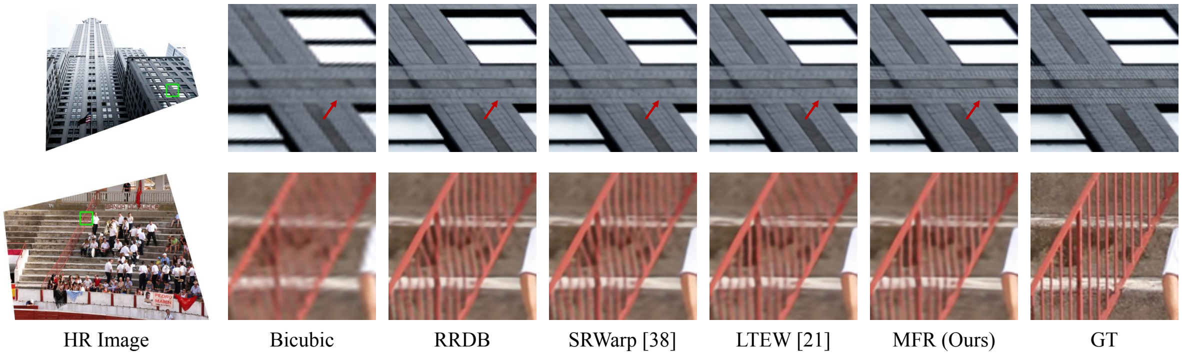

Figure 4. Illustration of the images generated by different image-warping methods in the in-scale setting.

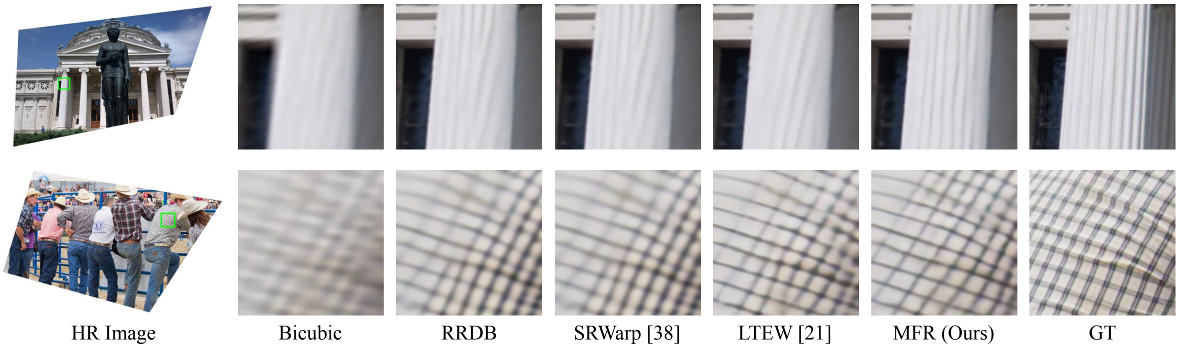

Figure 5. Illustration of the images generated by different image-warping methods in the out-of-scale setting.

compared models in all benchmark datasets. In particular, it outperforms the second-best model (i.e., LTEW) by 0.20dB and 0.26dB on the DIV2KW and Urban100W datasets, respectively. These results show that MFR has a superior generalization capability to handle out-of-scale images. Additionally, for visual comparison, we select two generated images from each evaluation setting in Fig. 4 and Fig. 5. As observed, SRWarp and LTEW have limited ability to generate texture and detailed information, resulting in distorted image content. In contrast, MFR can effectively synthesize high-frequency information, such as edges, textures, etc., leading to the generated images with high visual quality.

4.3. Experiments on ERP to Perspective Projection

In addition to evaluating the performance of MFR on images with out-of-distribution scales, we further explore its generalization capability on out-of-distribution transformations, i.e., perspective projection of ERP images. To conduct this investigation, we employ MFR, initially trained for homography transformation, on ERP images from the Flick360 validation dataset [7]. The size of the input ERP images is 2048 × 1024, and we project the images to the

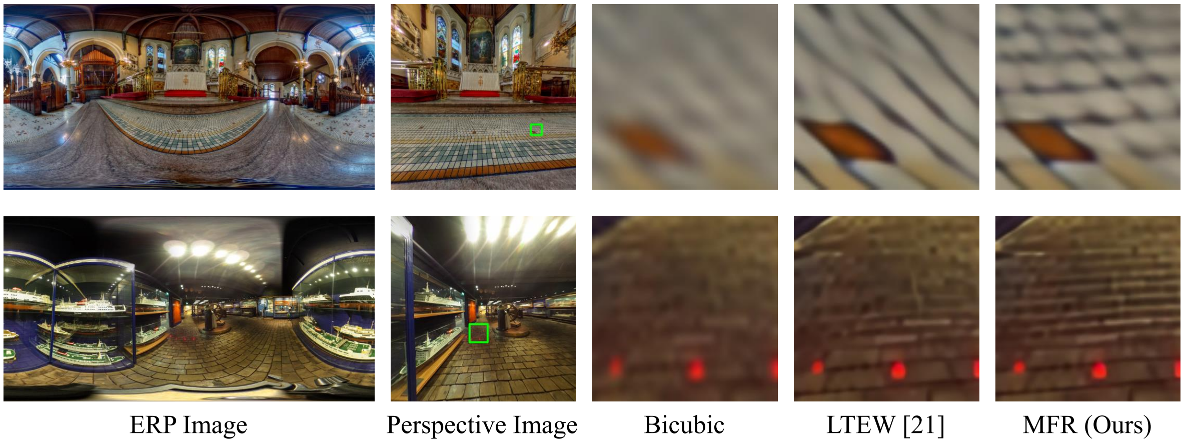

Figure 6. Illustration of the results (ERP images → Perspective images) generated by different image-warping methods.

size of 1024×1024 with a field of view (FOV) of 120°. As this dataset does not provide ground-truth images in the perspective view, we compare the performance of different methods through visual results. We choose two generated images and illustrate them in Fig. 6. For better comparison, we cropped two local regions marked by green rectangles and enlarged them.

As observed, the original EPR images suffer from serious deformation, but our MFR can effectively align them in the perspective views, while preserving local image content, compared with other image-warping methods. MFR produces images with less distortion and delivers clear details, resulting in superior visual quality.

4.4. Experiments on Asymmetric Image Super-resolution

We compare the proposed MFR with MetaSR [15], ArbSR [44], and LTEW [22] for asymmetric image super-resolution, in both the in-scale and out-of-scale settings. All methods adopt the RCAN model [53] as the backbone for a fair comparison. Four benchmark datasets, including Set5, Set14, B100, and Urban100, are used for evaluation. For the original RCAN model, we first upsample the images with a scaling factor of 4, followed by the bicubic interpolation to achieve the desired size.

Tables 2 and 3 illustrate the average PSNR scores of different methods for asymmetric image SR on benchmark datasets in the in-scale and out-of-scale settings, respectively. In the in-scaling setting, MFR can achieve better performance than the compared methods. Similarly, in the out-of-scaling setting, our MFR exhibits superior generalization capability on different benchmark datasets, in particular, for the B100 and Urban100 datasets with large scaling factors. In addition, we show the images generated by different methods under the in-scale and out-of-scale settings in Fig. 7 and Fig. 8, respectively. These results demonstrate that the proposed MFR has a better ability to generate high-frequency information, including textures and edges, than the compared methods, resulting in the best visual quality.

4.5. Ablation Study

4.5.1 Experiments on Various Filters

The choice of filters plays a significant role in learning frequency representations in our model. In this experiment, we investigate the impact of different filters on the performance. To facilitate the evaluation, we compare the learnable Gabor wavelet filter with the SIREN filter [40] and the conventional Gabor filter. Instead of adaptively updating the filter parameters, we evaluate the performance of our model using static Gabor filters. To ensure a fair comparison, all models employ EDSR [27] as the backbone for feature encoding. Table 4 shows the average mPSNR of our model using different filters for homography transformation on the DIV2KW dataset.

We find that the impact of employing different filters in our model is marginal in the in-scale setting but it becomes substantial in the out-of-scale setting. We demonstrate the images generated using different filters in Fig. 9. From these results, we find that employing learnable Gabor wavelet filters can significantly enhance MFR's capability to produce high-frequency details while avoiding distortion, which benefits from the compactness property of Gabor wavelet filters in both the spatial and frequency domains. This property enables our proposed MFR to effectively capture local variations in deformable images.

4.5.2 Experiments on Network Structures

Moreover, we study how the structure of the network affects the performance of our MFR. In the generation of deformable images, we adopt a short connection to fuse the spatial representations from the input with the fre-

Table 2. The average PSNR(dB) of state-of-the-art methods for asymmetric-scale SR on the benchmark datasets in the in-scale setting. The best and the second-best results are highlighted in red and blue, respectively.

Methods	Set5			Set14			B100			Urban100		
	$\frac{\times 1.5}{\times 4.0}$	$\frac{\times 1.5}{\times 3.5}$	$\frac{\times 1.6}{\times 3.05}$	$\frac{\times 4.0}{\times 2.0}$	$\frac{\times 3.5}{\times 2.0}$	$\frac{\times 3.5}{\times 1.75}$	$\frac{\times 4.0}{\times 1.4}$	$\frac{\times 1.5}{\times 3.0}$	$\frac{\times 3.5}{\times 1.75}$	$\frac{\times 1.6}{\times 3.0}$	$\frac{\times 1.6}{\times 3.8}$	$\frac{\times 3.55}{\times 1.55}$
Bicubic	30.01	30.86	31.40	27.25	27.88	27.27	27.45	28.86	27.94	25.93	24.92	25.19
RCAN [53]	34.14	35.05	35.67	30.35	31.02	31.21	29.35	31.30	29.98	30.72	28.81	29.34
MetaSR-RCAN [15]	34.20	35.17	35.81	30.40	31.05	31.33	29.43	31.26	30.09	30.73	29.03	29.67
ArbSR-RCAN [44]	34.37	35.40	36.05	30.55	31.27	31.54	29.54	31.40	30.22	31.13	29.36	30.04
LTEW-RCAN [22]	34.45	35.46	36.12	30.57	31.21	31.55	29.62	31.40	30.24	31.25	29.57	30.21
MFR-RCAN (Ours)	34.48	35.49	36.13	30.66	31.33	31.63	29.65	31.42	30.26	31.33	29.65	30.29

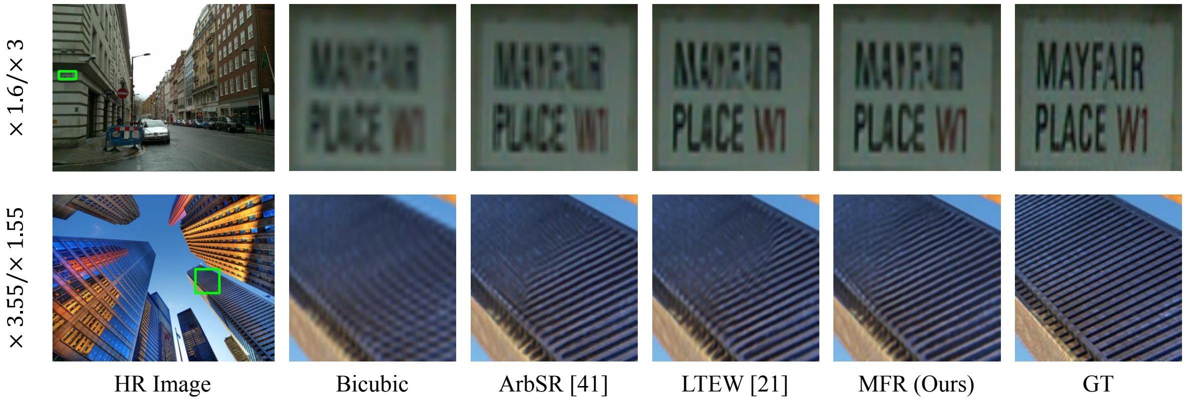

Figure 7. Illustration of the visual results generated by different asymmetric-scale SR methods in the in-scale setting.

Table 3. The average PSNR(dB) of state-of-the-art methods for asymmetric-scale SR on the benchmark datasets in the out-of-scale setting. The best and the second-best results are highlighted in red and blue, respectively.

Methods	Set5			Set14			B100			Urban100		
	$\frac{\times 3.0}{\times 8.0}$	$\frac{\times 3.0}{\times 7.0}$	$\frac{\times 3.2}{\times 6.1}$	$\frac{\times 8.0}{\times 4.0}$	$\frac{\times 7.0}{\times 4.0}$	$\frac{\times 7.0}{\times 3.5}$	$\frac{\times 8.0}{\times 2.8}$	$\frac{\times 3.0}{\times 6.0}$	$\frac{\times 7.0}{\times 2.9}$	$\frac{\times 3.2}{\times 6.0}$	$\frac{\times 3.2}{\times 7.6}$	$\frac{\times 7.1}{\times 3.1}$
Bicubic	25.69	26.35	26.84	24.27	24.62	24.79	24.67	25.58	24.98	22.55	21.92	22.15
RCAN [53]	29.00	30.01	30.46	26.48	26.94	27.11	26.06	27.19	26.47	25.52	24.50	24.84
MetaSR-RCAN [15]	28.75	29.74	30.38	26.32	26.85	27.03	26.07	27.15	26.45	25.50	24.47	24.84
ArbSR-RCAN [44]	28.37	29.35	30.08	26.06	26.63	26.84	25.91	27.14	26.40	25.36	24.12	24.61
LTEW-RCAN [22]	29.26	30.16	30.64	26.60	27.06	27.25	26.25	27.28	26.62	25.85	24.79	25.18
MFR-RCAN (Ours)	29.27	30.12	30.68	26.61	27.12	27.31	26.29	27.32	26.66	25.88	24.85	25.26

quency representations learned from the networks with a gate mechanism. To evaluate the effectiveness of these two mechanisms, we evaluate the model performance with and without using the short connection and the gate mechanism. The average mPSNR scores of MFR for homography transformation on the DIV2KW and Urban100W datasets are illustrated in Table 5.

Table 5 shows that MFR cannot achieve satisfactory performance by only using frequency representations, without the short connection. Using the gate mechanism, the perfor-

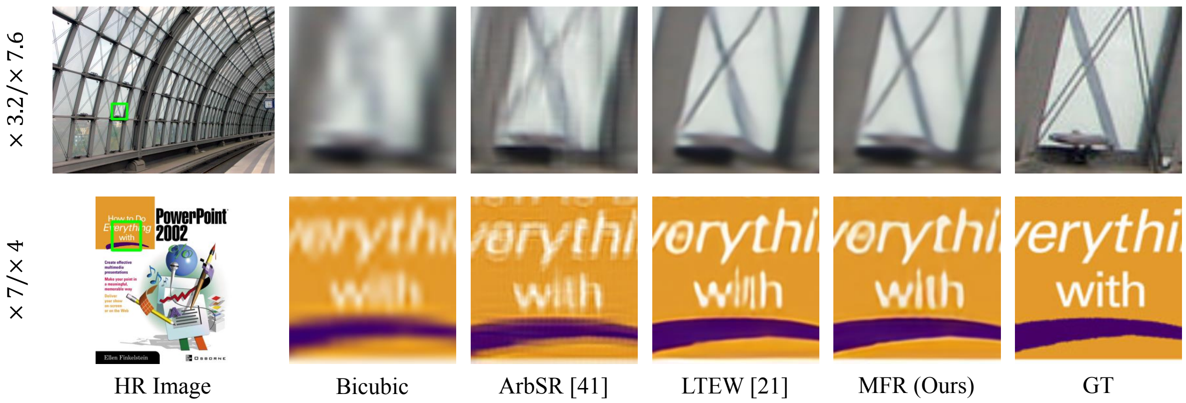

Figure 8. Illustration of the visual results generated by different asymmetric-scale SR methods in the out-of-scale setting.

Table 4. The average mPSNR of MFR using different filters for homography transformation on the DIV2K dataset. "GW-S" and "GW-L" denote the static and learnable Gabor wavelet filter, respectively. The best results are highlighted in bold.

	SIREN [40]	Gabor	GW-S	GW-L
is	30.87	30.83	30.86	**30.88**
os	26.81	26.86	26.86	**26.90**

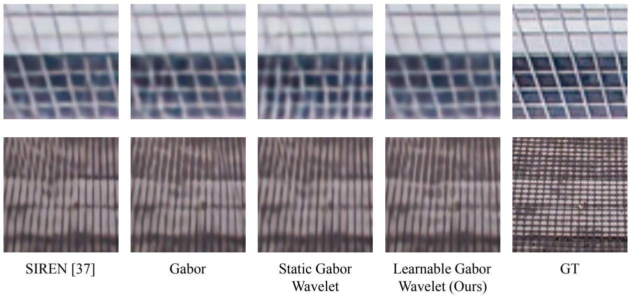

Figure 9. Illustration of visual results generated by our models using different filters.

Table 5. The average mPSNR of MFR using different network structures for homography transformation on the DIV2KW and Urban100W datasets. "SC" and "Gate" denote the short connection and the gate mechanism, respectively. The best results are highlighted in bold.

SC	Gate	DIV2KW		Urban100W	
		is	os	is	os
		30.72	26.80	29.02	24.07
✓		30.86	26.88	29.33	24.42
	✓	30.79	26.83	29.12	24.11
✓	✓	**30.88**	**26.90**	**29.68**	**24.51**

mance of the proposed MFR can be further improved. Overall, network structure is substantial for the performance of our MFR. However, it is worth noting that the optimal network structure should be tailored to specific applications and implementation scenarios.

5. Conclusion

In this paper, we propose a novel and effective method to learn the frequency representations of input images for image warping, namely MFR. Concretely, our MFR first employs a pretrained image super-resolution model to project the input image into the latent space. Then, we propose a filtering network to progressively learn frequency representations from different frequency subbands of the input features and generate deformable images in a coarse-to-fine manner. Furthermore, we incorporate Gabor wavelet filters into our model to enhance the capability to simultaneously capture local variations in deformable regions in both the spatial and frequency domains. Experiments show the superior performance of the proposed MFR in various tasks, including homography transformation, equirectangular to perspective projection, and asymmetry image super-resolution, significantly outperforming state-of-the-art image-warping methods. In addition, our MFR exhibits better generalization ability when processing out-of-distribution images with large scaling factors and transformations. The images generated by our model have rich detailed information and reduced distortion, resulting in the best visual quality.

Acknowledgments

This work was supported by the Hong Kong Research Grants Council (RGC) Research Impact Fund (RIF) under Grant R5001-18.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4432–4441, 2019. 1, 2

[2] 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. 4

[3] Per Rønsholt Andresen and Mads Nielsen. Non-rigid registration by geometry-constrained diffusion. *Medical Image Analysis*, 5(2):81–88, 2001. 2

[4] Reza Azad, Abdur R Fayjie, Claude Kauffmann, Ismail Ben Ayed, Marco Pedersoli, and Jose Dolz. On the texture bias for few-shot cnn segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2674–2683, 2021. 2

[5] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61:139–157, 2005. 2

[6] Thaddeus Beier and Shawn Neely. Feature-based image metamorphosis. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 529–536. 2023. 2

[7] Mingdeng Cao, Chong Mou, Fanghua Yu, Xintao Wang, Yinqiang Zheng, Jian Zhang, Chao Dong, Gen Li, Ying Shan, Radu Timofte, et al. Ntire 2023 challenge on 360deg omnidirectional image and video super-resolution: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1731–1745, 2023. 5

[8] Hao Cheng, Siyuan Yang, Joey Tianyi Zhou, Lanqing Guo, and Bihan Wen. Frequency guidance matters in few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11814–11824, 2023. 2

[9] Ming-Chao Chiang. *Imaging-consistent warping and super-resolution*. Columbia University, 1998. 1

[10] Rizal Fathony, Anit Kumar Sahu, Devin Willmott, and J Zico Kolter. Multiplicative filter networks. In *International Conference on Learning Representations*, 2020. 2

[11] Julia Gong, Yannick Hold-Geoffroy, and Jingwan Lu. Autotoon: Automatic geometric warping for face cartoon generation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 360–369, 2020. 1, 2

[12] Steven Haker, Lei Zhu, Allen Tannenbaum, and Sigurd Angenent. Optimal mass transport for registration and warping. *International Journal of computer vision*, 60:225–240, 2004. 1, 2

[13] Mark Holden. A review of geometric transformations for nonrigid body registration. *IEEE transactions on medical imaging*, 27(1):111–128, 2007. 2

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[15] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1575–1584, 2019. 6, 7

[16] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021. 2

[17] Muwei Jian, Kin-Man Lam, Junyu Dong, and Linlin Shen. Visual-patch-attention-aware saliency detection. *IEEE transactions on cybernetics*, 45(8):1575–1586, 2014. 4

[18] 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. 2

[19] Yakun Ju, Kin-Man Lam, Jun Xiao, Cong Zhang, Cuixin Yang, and Junyu Dong. Efficient feature fusion for learning-based photometric stereo. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[20] Verena Kaynig, Bernd Fischer, and Joachim M Buhmann. Probabilistic image registration and anomaly detection by nonlinear warping. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 1, 2

[21] Daniel Keysers, Thomas Deselaers, Christian Gollan, and Hermann Ney. Deformation models for image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1422–1435, 2007. 2

[22] Jaewon Lee, Kwang Pyo Choi, and Kyong Hwan Jin. Learning local implicit fourier representation for image warping. In *European Conference on Computer Vision*, pages 182–200. Springer, 2022. 1, 2, 4, 5, 6, 7

[23] Tai Sing Lee. Image representation using 2d gabor wavelets. *IEEE Transactions on pattern analysis and machine intelligence*, 18(10):959–971, 1996. 4

[24] Dong Li, Huiling Zhou, and Kin-Man Lam. High-resolution face verification using pore-scale facial features. *IEEE transactions on image processing*, 24(8):2317–2327, 2015. 4

[25] Jiacheng Li, Chang Chen, Wei Huang, Zhiqiang Lang, Fenglong Song, Youliang Yan, and Zhiwei Xiong. Learning steerable function for efficient image resampling. In *CVPR*, 2023. 2

[26] Zhenxuan Li, Wenzhong Shi, Hua Zhang, and Ming Hao. Change detection based on gabor wavelet features for very high resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 14(5):783–787, 2017. 4

[27] 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. 6

[28] David B Lindell, Dave Van Veen, Jeong Joon Park, and Gordon Wetzstein. Bacon: Band-limited coordinate networks for multiscale scene representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16252–16262, 2022. 2

[29] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical

image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021. 2

[30] Jan Modersitzki. *FAIR: flexible algorithms for image registration*. SIAM, 2009. 1

[31] Tony CW Mok and Albert CS Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 211–221. Springer, 2020. 1, 2

[32] Kuong-Hon Pong and Kin-Man Lam. Multi-resolution feature fusion for face recognition. *Pattern Recognition*, 47(2):556–567, 2014. 4

[33] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020. 2

[34] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 783–792, 2021.

[35] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in neural information processing systems*, 34:980–993, 2021. 2

[36] Enrico Segre. Image warp preserving content intensity. *SIAM Journal on Imaging Sciences*, 15(4):1623–1645, 2022. 1

[37] Shayan Shekarforoush, David Lindell, David J Fleet, and Marcus A Brubaker. Residual multiplicative filter networks for multiscale reconstruction. *Advances in Neural Information Processing Systems*, 35:8550–8563, 2022. 2

[38] Linlin Shen and Sen Jia. Three-dimensional gabor wavelets for pixel-based hyperspectral imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 49(12):5039–5046, 2011. 4

[39] Changjian Shui, Ruizhi Pu, Gezheng Xu, Jun Wen, Fan Zhou, Christian Gagné, Charles X Ling, and Boyu Wang. Towards more general loss and setting in unsupervised domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 2023. 2

[40] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 2, 3, 4, 6, 8

[41] Sanghyun Son and Kyoung Mu Lee. Srwarp: Generalized image super-resolution under arbitrary transformation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7782–7791, 2021. 1, 2, 4, 5

[42] 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. 2

[43] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8684–8694, 2020. 2

[44] Longguang Wang, Yingqian Wang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning a single network for scale-arbitrary super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4801–4810, 2021. 6, 7

[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. 4, 5

[46] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018. 1, 2

[47] Jun Xiao, Tianshan Liu, Rui Zhao, and Kin-Man Lam. Balanced distortion and perception in single-image super-resolution based on optimal transport in wavelet domain. *Neurocomputing*, 464:408–420, 2021. 2

[48] Jun Xiao, Xinyang Jiang, Ningxin Zheng, Huan Yang, Yifan Yang, Yuqing Yang, Dongsheng Li, and Kin-Man Lam. Online video super-resolution with convolutional kernel bypass grafts. *IEEE Transactions on Multimedia*, 2023. 2

[49] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1740–1749, 2020. 2

[50] 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

[51] Gizem Yüce, Guillermo Ortiz-Jiménez, Beril Besbinar, and Pascal Frossard. A structured dictionary perspective on implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19228–19238, 2022. 3

[52] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 1, 2

[53] 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. 6, 7