Towards Efficient and Scale-Robust Ultra-High-Definition Image Demoiréing

Xin Yu[1], Peng Dai[1], Wenbo Li[2], Lan Ma[3], Jiajun Shen[3], Jia Li[4], and Xiaojuan Qi[1✉]

[1] The University of Hong Kong
[2] The Chinese University of Hong Kong
[3] TCL AI Lab
[4] Sun Yat-sen University

Abstract. With the rapid development of mobile devices, modern widely-used mobile phones typically allow users to capture 4K resolution (i.e., ultra-high-definition) images. However, for image demoiréing, a challenging task in low-level vision, existing works are generally carried out on low-resolution or synthetic images. Hence, the effectiveness of these methods on 4K resolution images is still unknown. In this paper, we explore moiré pattern removal for ultra-high-definition images. To this end, we propose the first ultra-high-definition demoiréing dataset (UHDM), which contains 5,000 real-world 4K resolution image pairs, and conduct a benchmark study on current state-of-the-art methods. Further, we present an efficient baseline model ESDNet for tackling 4K moiré images, wherein we build a semantic-aligned scale-aware module to address the scale variation of moiré patterns. Extensive experiments manifest the effectiveness of our approach, which outperforms state-of-the-art methods by a large margin while being much more lightweight. Code and dataset are available at https://xinyu-andy.github.io/uhdm-page.

Keywords: Image demoiréing, Image restoration, Ultra-high-definition

1 Introduction

When photographing the contents displayed on the digital screen, an inevitable frequency aliasing between the camera's color filter array (CFA) and the screen's LCD subpixel widely exists. The captured images are thus mixed with colorful stripes, named moiré patterns, which severely degrade the perceptual quality of images. Currently, efficiently removing moiré patterns from a single moiré image is still challenging and receives growing attention from the research community.

Recently, several image demoiréing methods [13,46,12,29,22,8,20,40] have been proposed, yielding a plethora of dedicated designs such as moiré pattern classification [12], frequency domain modeling [22,46], and multi-stage framework [13]. Apart from FHDe^{2}Net [13] which is specially designed for high-definition images, most of the research efforts have been devoted to studying low-resolution images

✉ indicates the corresponding author.

[29] or synthetic images [40]. However, the fast development of mobile devices enables modern mobile phones to capture ultra-high-definition images, so it is more practical to conduct research on 4K image demoiréing for real applications. Unfortunately, the highest resolution in current public demoiréing datasets (see Table 1) is 1080p [13] (1920×1080). Whether methods investigated on such datasets can be trivially transferred into the 4K scenario is still unknown due to the data distribution change and dramatically increased computational cost.

Under this circumstance, we explore the more practical yet more challenging demoiréing scenario, i.e., ultra-high-definition image demoiréing. To evaluate the demoiréing methods in this scenario, we build the first large-scale real-world ultra-high-definition demoiréing dataset (UHDM), which consists of $4,500$ training image pairs and 500 testing image pairs with diverse scenes (see Fig. 1).

Benchmark study and limitation analysis: Based upon our dataset, we conduct a benchmark study on state-of-the-art methods [13,46,12,29,22,8]. Our empirical study reveals that most methods [29,8,46] struggle to remove moiré patterns with a much wider range of scales in 4K images while simultaneously tolerating the growing demands for computational cost (see Fig. 3) or fine image detail [13] (see Fig. 2). We attribute their deficiencies to the lack of an effective multi-scale feature extraction strategy. Concretely, existing methods attempting to address the scale challenge can be coarsely categorized into two lines of research. One line of research develops multi-stage models, such as FHDe^{2}Net [13], to process large moiré patterns at a low-resolution stage and then refines the textures at a high-resolution stage, which however incurs huge computational cost when applied to 4K images (see Fig. 3: FHDe^{2}Net). Another line of research utilizes features from different depths of a network to build multi-scale representations, in which the most representative work [46] achieves a better trade-off between accuracy and efficiency (see Fig. 3: MBCNN), yet still cannot be generally scale-robust (see Fig. 2 and Fig. 5). We note that the extracted multi-scale features are from different semantic levels which may result in misaligned features when fused together, potentially limiting its capabilities. Detailed study and analysis are unfolded in Section 3.2.

To this end, inspired by HRNet [33], we propose a plug-and-play semantic-aligned scale-aware module (SAM) to boost the network's capability in handling moiré patterns with diverse scales without incurring too much computational cost, serving as a supplement to existing methods. Specifically, SAM incorporates a pyramid context extraction module to effectively and efficiently extract multi-scale features aligned at the same semantic level. Further, a cross-scale dynamic fusion module is developed to selectively fuse multi-scale features where the fusion weights are learned and dynamically adapted to individual images.

Equipped with SAM, we develop an efficient and scale-robust network for 4K image demoiréing, named ESDNet. ESDNet adopts a simple encoder-decoder network with skip-connections as its backbone and stacks SAM at different semantic levels to boost the model's capability in addressing scale variations of 4K moiré images. ESDNet is easy to implement while achieving state-of-the-art performance (see Fig. 5 and Table 2) on the challenging ultra-high-definition image demoiréing dataset and three other public demoiréing datasets [13,40,29]. In particular, ESDNet exceeds multi-stage high-resolution method FHDe^{2}Net, **1.8dB** in terms of PSNR while being **300× faster (5.620s vs 0.017s)** in the UHDM dataset. Our major contributions are summarized as follows:

- We are the first to explore the ultra-high-definition image demoiréing problem, which is more practical yet more challenging. To this end, we build a large-scale real-world 4K resolution demoiréing dataset UHDM.
- We conduct a benchmark study for the existing state-of-the-art methods on this dataset, summarizing several challenges and analyses. Motivated by these analyses, we propose an efficient baseline model ESDNet for ultra-high-definition image demoiréing.
- Our ESDNet achieves state-of-the-art results on the UHDM dataset and three other public demoiréing datasets, in terms of quantitative evaluation and qualitative comparisons. Moreover, ESDNet is lightweight and can process standard 4K (3840 × 2160) resolution images at 60 fps.

2 Related Work

Image demoiréing: To remove moiré patterns caused by the frequency aliasing, Liu et al. [20] propose a synthetic dataset by simulating the camera imaging process and develop a GAN-based [10] framework. Further, a large-scale synthetic dataset [40] is proposed and promotes many follow-up works [46,8,40]. However, it is difficult for models trained on synthetic data to handle real-world scenarios due to the sim-to-real gap. For real-world image demoiréing, Sun et al. [29] propose the first real-world moiré image dataset (i.e., TIP2018) and develop a multi-scale network (DMCNN). To distinguish different types of moiré patterns, He et al. [12] manually annotate moiré images with category labels to train a moiré pattern classification model. Frequency domain methods [22,46] are also studied for moiré removal. To deal with high-resolution images, He et al. [13] construct a high-definition dataset FHDMi and develop the multi-stage framework FHDe^{2}Net. Although significant progress has been achieved, the above methods either cannot achieve satisfactory results [46,12,29,8] or suffer from heavy computational cost [46,13,12,8]. More importantly, the highest resolution of existing image demoiréing datasets is FHDMi [13] with 1080p resolution, which is not suitable for practical use considering the ultra-high-definition (4K) images captured by current mobile cameras. We focus on developing a lightweight model that can process ultra-high-definition images.

Image restoration: To this point, plenty of learning-based image restoration models have been proposed. For instance, residual learning [14] and dense connection [15] are widely used to develop very deep neural networks for different low-level vision tasks [43,1,19,17,45]. In order to capture multi-scale information, encoder-decoder [25] structures or hierarchical architectures are frequently exploited in image restoration tasks [42,41,9]. Inspired by iterative solvers, some methods utilize recurrent structures [9,31] to gradually recover images while reducing the number of parameters. To preserve structural and semantic information, many works [36,21,28,37,30,34] adopt the perceptual loss [16] or generative loss [10,11,2] to guide the training procedure. In our work, we also take advantage of the well-designed dense blocks for efficient feature reuse and the perceptual loss for semantically guided optimization.

Multi-scale network: The multi-scale network has been widely adopted in various tasks [33,4,47,38,6] due to its ability to leverage features with different receptive fields. U-Net [25], as one representative multi-scale network, extracts multi-scale information using an encoder-decoder structure, and enhances features in decoder with skip-connections. To preserve the high-resolution representation, the full resolution residual network [24] extends the U-Net by introducing an extra stream containing information of the full resolution, and similar operations can be found in the HRNet [33]. Considering that the extracted multi-scale features have different semantic meanings, the question of how to fuse features with different meanings is also important and has been widely studied in many works [3,5,7]. In this work, we design a semantic-aligned scale-aware module to handle moiré patterns with diverse scales without incurring too great a computational cost, which renders our method highly practical for 4K images.

3 UHDM Dataset

We study ultra-high-definition image demoiréing, which has more practical applications. For the training of 4K demoiréing models and the evaluation of existing methods, we collect a large-scale ultra-high-definition demoiréing dataset (UHDM). Dataset collection and benchmark study are elaborated upon below.

3.1 Data Collection and Selection

To obtain the real-world 4K image pairs, we first collect high-quality images with resolutions ranging from 4K to 8K from the Internet. We note that Internet resources lack document scenes, which also constitute a vital application scenario (e.g., slides, papers), so we manually generate high-quality text images and make sure they maintain 3000 dpi (Dots Per Inch). Finally, the collected moiré-free images cover a wide range of scenes (see Fig. 1), such as landscapes, sports, video clips, and documents. Given these high-quality images, we generate diverse real-world moiré patterns elaborated upon below.

First, to produce realistic moiré images and ease the difficulties of calibrations, we shoot the clean pictures displayed on the screen with a camera phone fixed on a DJI OM 5 smartphone gimbal, which allows us to conveniently and flexibly adjust the camera view through its control button, as shown in Fig. 1. Second, we note that the characteristics of moiré patterns highly are highly dependent upon the geometric relationship between the screen and the camera (see

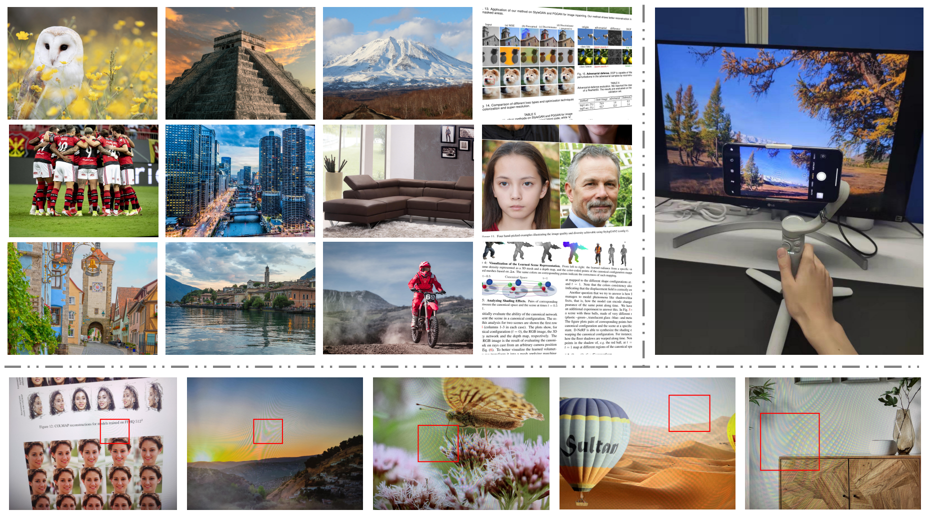

Fig. 1: Upper left: Our dataset contains diversified scenarios. Upper right: we capture the moiré image with a DJI OM 5 smartphone gimbal. Lower: moiré images in our dataset show a wide range of scale variations

supplement for more details). Therefore, during the capturing process, we continuously adjust the viewpoint every ten shots to produce diverse moiré patterns. Third, we adopt multiple $<$ mobile phone, screen $>$ (i.e., three mobile phones and three digital screens, see supplement for more details) combinations to cover various device pairs, since they will also have an impact on the styles of moiré patterns. Finally, to obtain aligned pairs, we utilize RANSAC algorithm [32] to estimate the homography matrix between the original high-quality image and the captured moiré screen image. Since it is difficult to ensure accurate pixel-wise calibration due to the camera's internal nonlinear distortions and perturbations of moiré artifacts, manual selection is performed to rule out severely misaligned image pairs, thereby ensuring quality.

Our dataset contains $5,000$ image pairs in total. We randomly split them into $4,500$ for training and 500 for validation. As we collect moiré images using various mobile phones, the resolution can either be 4032×3024 or 4624×3472. Comparisons with other existing datasets are shown in Table 1, and the characteristics of our dataset are summarized as below.

- **Ultra-high resolution** UHDM is the first 4K resolution demoiréing dataset, consisting of 5,000 image pairs in total.
- **Diverse image scenes** The dataset includes diverse scenes, such as landscapes, sports, video clips, and documents.
- **Real-world capture settings** The moiré images are generated following practical routines, with different device combinations and viewpoints to produce diverse moiré patterns.

Table 1: Comparisons of different demoiréing datasets; our dataset is the first ultra-high-definition dataset ("London's Buildings" is not available currently)

Dataset	Avg. Resolution	Size	Diversity	Real-world
TIP18 [29]	256×256	135,000	No text scenes	✓
LCDMoiré [40]	1024×1024	10,200	Only text scenes	×
FHDMi [13]	1920×1080	12,000	Diverse scenes	✓
London's Buildings [22]	2100×1700	460	Only urban scenes	✓
UHDM	$\mathbf{4328 \times 3248}$	**5,000**	**Diverse scenes**	✓

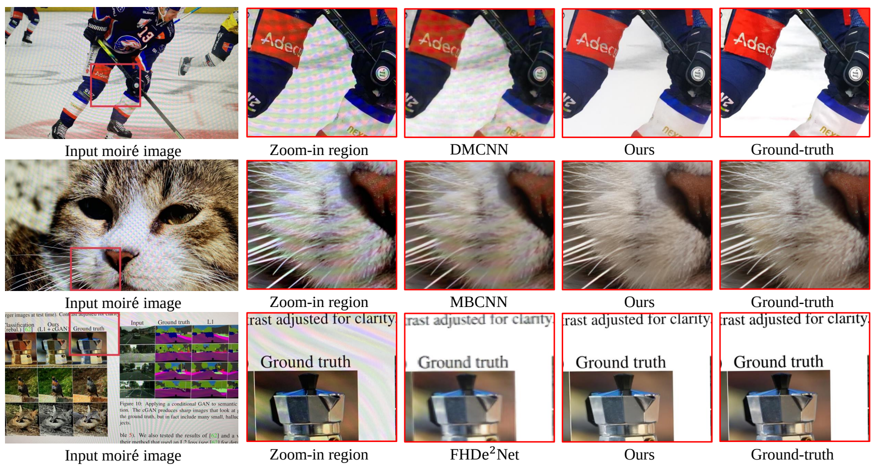

Fig. 2: Limitations of current methods: they are often unable to remove the moiré pattern with a wider scale range or lose high-frequency details

3.2 Benchmark Study on 4K Demoiréing

As the image resolution is increased to the 4K resolution, the scale of moiré patterns has a very wide range, from very large moiré patterns to small ones (see Fig. 1). This poses a major challenge to demoiréing methods as they are required to be scale-robust. Furthermore, increased image resolution also leads to dramatically increased computational cost and high requirements of detail restoration/preservation. Here, we carry out a benchmark study on the existing state-of-the-art methods [46,29,12,13,22,8] on our 4K demoiréing dataset to evaluate their effectiveness. Main results are summarized in Fig. 2 and Fig. 3: existing methods are mostly not capable of achieving a good balance of accuracy and computational efficiency. More detailed results are shown in Section 5.

Analysis and discussions: Although existing methods also attempt to address the scale challenge by developing multi-scale strategies, they still have several deficiencies regarding computational efficiency and restoration quality when applied to 4K high-resolution images (see Fig. 2). One line of methods, such as DMCNN [29] and MDDM [8], fuses multi-scale features harvested from multi-resolution inputs only at the output stage, which potentially prohibits

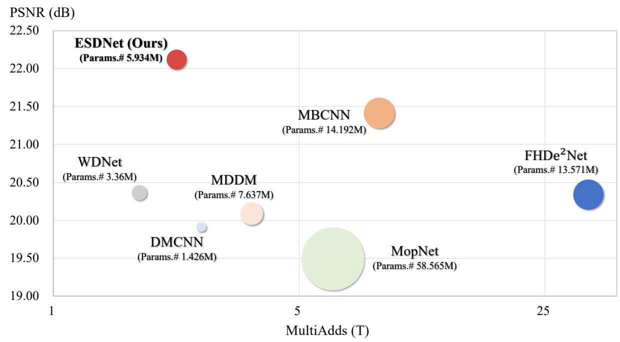

Fig. 3: Comparisons of computational cost of different methods: The x-axis and the y-axis denote the MultiAdds (T) and PSNR (dB). The number of parameters is expressed by the area of the circle

the intermediate features from interacting with and refining each other, leading to sub-optimal results, i.e., significantly sacrificing accuracy on 4K image demoiréing despite being lightweight (see Fig. 3 and Fig. 2). Another line of methods, such as MBCNN [46], exploits multi-scale features at different network depths following a U-Net-like architecture. Compared with other existing methods, although it achieves the best trade-off between accuracy and efficiency, it still suffers from moiré patterns with a wide-scale range (the second row of Fig. 2 and Fig. 5). One possible issue is that the combined multi-scale features come from different semantic levels [33], prohibiting a specific feature level to harvest multi-resolution representations [33], which could also be an important cue for image demoiréing. On the other hand, FHDe^{2}Net [13] designs a coarse-to-fine two-stage model to simultaneously address the scale and detail challenge. It suffers, however, from heavy computational cost when applied to 4K images (see Fig. 3) yet is still not sufficient to remove moiré patterns (see Fig. 5) or recover fine image detail (see Fig. 2 and Fig. 5).

4 Proposed Method

Motivated by observations in Section 3.2, we introduce a baseline approach to advance 4K resolution image demoiréing, aimed towards a more scale-robust and efficient model. In the following, we first present an overview of our pipeline and then elaborate on our core semantic-aligned scale-aware module (SAM).

4.1 Pipeline

The overall architecture is shown in Fig. 4, where a pre-processing head is utilized to enlarge the receptive field, followed by an encoder-decoder architecture

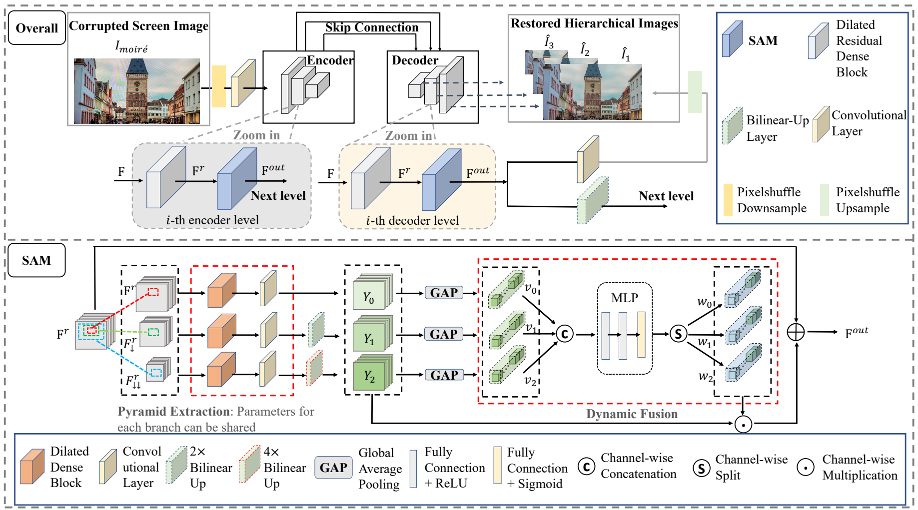

Fig. 4: The pipeline of our ESDNet and the proposed semantic-aligned scale-aware module (SAM)

for image demoiréing. The pre-processing head adopts pixel shuffle [26] to downsample the image by two times and a 5×5 convolution layer to further extract low-level features. Then, the extracted low-level features are fed into an encoder-decoder backbone architecture that consists of three downsampling and upsampling levels. Note that the encoder and decoder are connected via skip-connections to allow features containing high-resolution information to facilitate the restoration of corresponding moiré-free images. At each decoder level, the network would produce intermediate results through a convolution layer and a pixelshuffle upsampling operation (see the upper part of Fig. 4), which are also supervised by the ground-truth, serving the purpose of deep supervision to facilitate training. Specifically, each encoder or decoder level (see Fig. 4) contains a dilated residual dense block [45,15,14,39] for refining the input features (as detailed below) and a proposed semantic-aligned multi-scale module (SAM) for extracting and dynamically fusing multi-scale features at the same semantic level (as elaborated in Section 4.2).

Dilated residual dense block: For each level $i \in \{1, 2, 3, 4, 5, 6\}$ (i.e., three encoder levels and three decoder levels), the input feature F_i first goes through a convolutional block, i.e., dilated residual dense block, for refining input features. It incorporates the residual dense block (RDB) [45,15,14] and dilated convolution layers [39] to process the input features and output refined ones. Specifically, given an input feature F_i^0 to the i-th level encoder or decoder, the cascaded local features from each layer inside the block can be formulated as Eq. (1):

$$F_i^l = C^l([F_i^0, F_i^1, ..., F_i^{l-1}]), (l = 1, 2, ..., L) \tag{1}$$

where $[F_i^0, F_i^1, ..., F_i^{l-1}]$ denotes the concatenation of all intermediate features inside the block before layer l, and C^l is the operator to process the concatenated features, consisting of a 3×3 Conv with dilated rate d^l and a rectified linear unit (ReLU). After that, we apply a 1×1 convolution to keep the output channel number the same as that of F_i^0. Finally, we exploit the residual connection to produce the refined feature representation F_i^r, formulated as Eq.(2):

$$F_i^r = F_i^0 + \mathrm{Conv}_{1\times1}(F_i^L). \tag{2}$$

The refined feature representation F_i^r is then fed to our proposed SAM for semantic-aligned multi-scale feature extraction.

4.2 Semantic-Aligned Scale-Aware Module

Given the input feature F_i^r, the SAM is intended to extract multi-scale features within the same semantic level i and allow them to interact and be dynamically fused, significantly improving the model's ability to handle moiré patterns with a wide range of scales. As demonstrated in Table 3, SAM enables us to develop a lightweight network while still being more effective in comparison with existing methods. In the following, we detail the design of SAM which encompasses two major modules: pyramid feature extraction and cross-scale dynamic fusion.

Pyramid context extraction: Given an input feature map $F^r \in \mathbb{R}^{H\times W\times C}$ (we simplify F_i^r by F^r in the following discussion), we first produce pyramid input features $F^r \in \mathbb{R}^{H\times W\times C}, F_{\downarrow}^r \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$ and $F_{\downarrow\downarrow}^r \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$ through bilinear interpolation, then feed them into a corresponding convolutional branch with five convolution layers to yield pyramid outputs Y_0, Y_1, Y_2 (see the lower part of Fig. 4):

$$Y_0 = E_0(F^r), \quad Y_1 = E_1\left(F_{\downarrow}^r\right), \quad Y_2 = E_2\left(F_{\downarrow\downarrow}^r\right), \tag{3}$$

where we build E_0, E_1, and E_2 via the dilated dense block, followed by a 1×1 convolution layer. In addition, the up-sampling operations will be performed in E_1, E_2 to align the size of three outputs, i.e., $Y_i \in \mathbb{R}^{H\times W\times C}, (i = 0, 1, 2)$.

Note that, as the internal architectures of E_0, E_1, and E_2 are identical, their corresponding learnable parameters can be shared to lower the cost of parameter number. In fact, as proven in Section 5, the improvement primarily comes from the pyramid architecture instead of additional parameters.

Cross-scale dynamic fusion: Given the pyramid features Y_0, Y_1, Y_2, the cross-scale dynamic fusion module fuses them together to produce fused multi-scale features for the next level to process. The insight for this module is that scale of moiré patterns vary from image to image and thus the importance of different scale features would also vary across images. Therefore, we develop the following cross-scale dynamic fusion module to make the fusion process dynamically adjusted and adapted to each image. Specifically, we learn dynamic weights to fuse Y_1, Y_2, Y_3.

Given $Y_i \in \mathbb{R}^{H \times W \times C} (i = 0, 1, 2)$, we first apply global average pooling in the spatial dimension of each feature map to obtain the 1D global feature $v_i \in \mathbb{R}^C$ for each scale i following Eq. (4).

$$v_i = \frac{1}{H \times W} \sum_{s=1}^{H} \sum_{t=1}^{W} Y_i(s, t) \tag{4}$$

Then, we concatenate them along the channel dimension and learn the dynamic weights through an MLP module as:

$$[w_0, w_1, w_2] = \mathrm{MLP}([v_0, v_1, v_2]) \tag{5}$$

where "MLP" consists of three fully connected layers and outputs $w_0, w_1, w_2 \in \mathbb{R}^C$ to fuse Y_1, Y_2, Y_3 dynamically. Finally, with fusion weights, we channel-wisely fuse the pyramid features with the input-adaptive weights, and then add the input feature F^r to get the final output of SAM:

$$F^{\mathrm{out}} = F^r + w_0 \odot Y_0 + w_1 \odot Y_1 + w_2 \odot Y_2 \tag{6}$$

where $\odot$ denotes the channel-wise multiplication, and the output F^{out} will go through the next level ($i \rightarrow i + 1$) for further feature extraction and image reconstruction.

Comparisons and analysis: Existing methods [46,22] utilize features from different depths to obtain multi-scale representations. However, features at different depths have different levels of semantic information. Thus, they are incapable of representing multi-scale information at the same semantic level, which might provide important cues for boosting the model's multi-scale modeling capabilities, as indicated in [33]. We offer SAM as a supplement to existing methods as Y_0, Y_1, Y_2 include semantic-aligned information with different local receptive fields. The dynamic fusion methods further make the module adaptive to different images and boost its abilities. This strategy can also be treated as an implicit classifier compared with the explicit one in MopNet [12], which is more efficient and avoids the ambiguous hand-craft attribute definition. We include more detailed analysis in our supplementary file.

4.3 Loss Function

To boost optimization, we adopt the deep supervision strategy, which has been proven useful in [46]. As shown in Fig. 4, in each decoder level, the network will produce hierarchical predictions $\hat{I}_1, \hat{I}_2, \hat{I}_3$, which are also supervised by ground-truth images. We note that moiré patterns disrupt image structures since they generate new strip-shaped structures. Therefore, we adopt the perceptual loss [16] for feature-based supervision. At each level, we build our loss function

Table 2: Quantitative comparisons between our model and state-of-the-art methods on four datasets. (↑) denotes the larger the better, and (↓) denotes the smaller the better. Red: best and Blue: second-best

Dataset	Metrics	Input	DMCNN[29]	MDDM[8]	WDNet[22]	MopNet[12]	MBCNN[46]	FHDe^{2}Net[13]	ESDNet	ESDNet-L
UHDM	PSNR↑	17.117	19.914	20.088	20.364	19.489	21.414	20.338	22.119	22.422
	SSIM↑	0.5089	0.7575	0.7441	0.6497	0.7572	0.7932	0.7496	0.7956	0.7985
	LPIPS↓	0.5314	0.3764	0.3409	0.4882	0.3857	0.3318	0.3519	0.2551	0.2454
FHDMi	PSNR↑	17.974	21.538	20.831	-	22.756	22.309	22.930	24.500	24.882
	SSIM↑	0.7033	0.7727	0.7343	-	0.7958	0.8095	0.7885	0.8351	0.8440
	LPIPS↓	0.2837	0.2477	0.2515	-	0.1794	0.1980	0.1688	0.1354	0.1301
TIP2018	PSNR↑	20.30	26.77	-	28.08	27.75	30.03	27.78	29.81	30.11
	SSIM↑	0.738	0.871	-	0.904	0.895	0.893	0.896	0.916	0.920
LCDMoiré	PSNR↑	10.44	35.48	42.49	29.66	-	44.04	41.40	44.83	45.34
	SSIM↑	0.5717	0.9785	0.9940	0.9670	-	0.9948	-	0.9963	0.9966
-	Params (M)	-	1.426	7.637	3.360	58.565	14.192	13.571	5.934	10.623

by combining the pixel-wise L_1 loss and the feature-based perceptual loss L_p. Hence, the final loss function is formulated as:

$$\mathcal{L}_{total} = \sum_{i=1}^{3} \mathcal{L}_1(I_i, \hat{I}_i) + \lambda \times \mathcal{L}_p(I_i, \hat{I}_i) \tag{7}$$

For the perceptual loss, we extract features from conv3_3 (after ReLU) using a pre-trained VGG16 [27] network and compute the L_1 distance in the feature space; we simply set $\lambda = 1$ during training. We find that this perceptual loss is effective in removing moiré patterns.

5 Experiments

Datasets and metrics: We conduct experiments on the proposed UHDM dataset and three other public datasets: FHDMi [13], TIP2018 [29] and LCDMoiré [40]. In our UHDM dataset, we keep the original two resolutions (see Section 3) and models are trained with cropped patches. During the evaluation phase, we do center crop from the original images to obtain test pairs with a resolution of 3840×2160 (standard 4K size). We adopt the widely used PSNR, SSIM [35] and LPIPS [44] metrics for quantitative evaluation. It has been proven that LPIPS is more consistent with human perception and suitable for measuring demoiréing quality [13]. Note that existing methods only report PSNR and SSIM on the TIP2018 and LCDMoiré, so we follow this setup for comparisons.
Implementation details: We implement our algorithm using PyTorch on an NVIDIA RTX 3090 GPU card. During training, we randomly crop a 768×768 patch from the ultra-high-definition images, and set the batch size to 2. The model is trained for 150 epochs and optimized by Adam [18] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initially set to 0.0002 and scheduled by cyclic cosine annealing [23]. Details for implementations on other benchmarks are unfolded in the supplementary file. We also train other methods on our dataset faithfully and sufficiently and unfold details in the supplementary file.

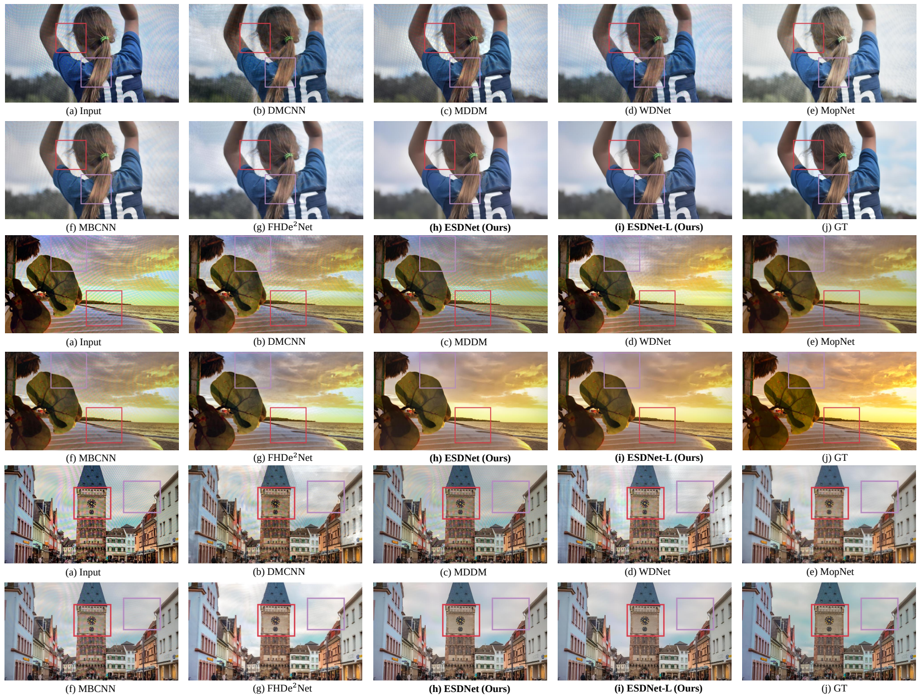

Fig. 5: Qualitative comparisons with state-of-the-art methods on the UHDM dataset. Please zoom in for a better view. More results are given in the supplementary file

5.1 Comparisons with State-of-the-Art Methods

We provide two versions of our model: ESDNet and ESDNet-L. ESDNet is the default lightweight model and ESDNet-L is a larger model, stacking one more SAM in each network level.

Quantitative comparison: Table 2 shows quantitative performance of existing approaches. The proposed method achieves state-of-the-art results on all four datasets. Specifically, both of our two models outperform other methods by a large margin in the ultra-high-definition UHDM dataset and high-definition FHDMi dataset, demonstrating the effectiveness of our method in high-resolution scenarios. It is worthwhile to note that our ESDNet, though possessing far fewer parameters, already shows competitive performance.

Qualitative comparison: We present visual comparisons between our algorithm and existing methods in Fig. 5. Apparently, our method obtains more perceptually satisfactory results. In comparison, MDDM [8], DMCNN [29] and WDNet [22] often fail to remove moiré patterns, while MBCNN [46] and MopNet [12] cannot handle large-scale patterns well. Though performing better than other methods (except for ours), FHDe^{2}Net [13] usually suffers from severe loss of details. All these facts manifest the superiority of our method.

Table 3: Ablation study of the proposed SAM. "A" represents the baseline model. "A^+" denotes a stronger baseline which is of similar model capacity compared to our full model "E". "B" adds the pyramid context extraction with shared weights across all branches to "A" while "D" adopts adaptive weights. "C" and "E" add the cross-scale dynamic fusion based on "B" and "D", respectively

Dataset	Metrics	A	A^+	B	C	D	E
	PSNR↑	20.646	20.860	21.176	21.958	21.300	**22.119**
UHDM	SSIM↑	0.7899	0.7908	0.7937	0.7938	0.7947	**0.7956**
	LPIPS↓	0.2750	0.2626	0.2683	0.2596	0.2623	**0.2551**
	Params (M)	2.705	5.978	2.705	3.014	5.625	5.934

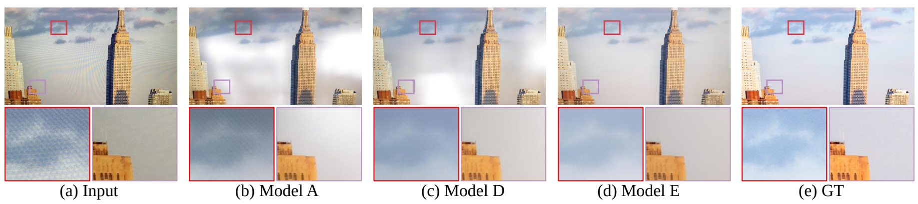

Fig. 6: Qualitative effects of different components in SAM

Computational cost: As shown in Fig. 3, our method strikes a sweet point of balancing the parameter number, computation cost (MACs), and demoiréing performance. Also, we test the inference speed of our method on an NVIDIA RTX 3090 GPU. Surprisingly, our ESDNet only needs 17ms (i.e., 60fps) to process a standard 4K resolution image, almost 300× faster than FHDe^2Net. The competitive performance and low computational cost render our method highly practical in the 4K scenario.

5.2 Ablation Study

In this section, we tease apart which components of our network contribute most to the final performance on the UHDM dataset. As shown in Table 3, we start from the baseline model (model "A"), which ablates the pyramid context extraction and the cross-scale dynamic fusion strategies. To make a fair comparison, we further build a stronger baseline model (model "A^+") that is comparable to our full model (model "E") in terms of the model capacity.

Pyramid context extraction: We construct two variants (model "B" and model "D") for exploring the effectiveness of this design. Compared with the baseline (model "A"), we observe that the proposed pyramid context extraction can significantly boost the model performance. To validate whether the improvement comes from more parameters in the additional two sub-branches, we exploit a weight-sharing strategy across all branches (model "B"). The observations in Table 3 demonstrate that the performance gain mainly stems from the pyramid design rather than the increase of parameters. Further, as shown in Fig. 6, we

Table 4: Ablation study of the loss function. The left and the right of "/" denote results trained by the pixel-wise L_1 loss and trained by our loss, respectively

Dataset	Metrics	DMCNN	MDDM	Ours
UHDM	PSNR↑	**19.914**/19.911	20.088/**20.333**	21.489/**22.119**
	SSIM↑	**0.7575**/0.7212	**0.7441**/0.7412	0.7893/**0.7956**
	LPIPS↓	0.3764/**0.3089**	0.3409/**0.2986**	0.3330/**0.2551**

find our pyramid design can successfully remove the moiré patterns that are not well addressed in the baseline model.

Cross-scale dynamic fusion: To verify the importance of the proposed dynamic fusion scheme, we increasingly add this design to model "B" and model "D", resulting in model "C" and model "E". We observe consistent improvements for both models, especially on PSNR. Also, Fig. 6 shows that the artifacts retained in model "D" are totally removed in the result of model "E", achieving a more harmonious color style.

Loss function: Through our experiments, we find the perceptual loss plays an essential role in image demoiréing. As shown in Table 4, when replacing our loss function with a single L_1 loss, we notice obvious performance drops in our method, especially on LPIPS. Also, we make further exploration by applying our loss function to other state-of-the-art methods [29,8]. The significant improvements on LPIPS illustrate the importance of the loss design in yielding a higher perceptual quality of recovered images. We suggest our loss is more robust to address the large-scale moiré patterns and the misaligned issue in the real-world datasets [13,29]. More discussions are included in the supplementary file.

6 Conclusion

In this paper, to explore the more practical yet challenging 4K image demoiréing scenario, we propose the first real-world ultra-high-definition demoiréing dataset (UHDM). Based upon this dataset, we conduct a benchmark study and limitation analysis of current methods, which motivates us to build a lightweight semantic-aligned scale-aware module (SAM) to strengthen the model's multi-scale ability without incurring much computational cost. By leveraging SAM in different depths of a simple encoder-decoder backbone network, we develop ESDNet to handle 4K high-resolution image demoiréing effectively. Our method is computationally efficient and easy to implement, achieving state-of-the-art results on four benchmark demoiréing datasets (including our UHDM). We hope our investigation could inspire future research in this more practical setting.

Acknowledgements. This work is partially supported by HKU-TCL Joint Research Center for Artificial Intelligence, Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27209621), National Key R&D Program of China (No.2021YFA1001300), and Guangdong-Hong Kong-Macau Applied Math Center grant 2020B1515310011.

References

1. Anwar, S., Barnes, N.: Densely residual laplacian super-resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 3
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017) 4
3. Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection. In: European conference on computer vision. pp. 354–370. Springer (2016) 4
4. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3291–3300 (2018) 4
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017) 4
6. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1511–1520 (2017) 4
7. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7103–7112 (2018) 4
8. Cheng, X., Fu, Z., Yang, J.: Multi-scale dynamic feature encoding network for image demoiréing. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 3486–3493. IEEE (2019) 1, 2, 3, 6, 11, 12, 14
9. Gao, H., Tao, X., Shen, X., Jia, J.: Dynamic scene deblurring with parameter selective sharing and nested skip connections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3848–3856 (2019) 3, 4
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014) 3, 4
11. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028 (2017) 4
12. He, B., Wang, C., Shi, B., Duan, L.Y.: Mop moire patterns using mopnet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2424–2432 (2019) 1, 2, 3, 6, 10, 11, 12
13. He, B., Wang, C., Shi, B., Duan, L.Y.: Fhde 2 net: Full high definition demoireing network. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 713–729. Springer (2020) 1, 2, 3, 6, 7, 11, 12, 14
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 3, 8
15. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017) 3, 8
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016) 4, 10

17. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016) 3
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 11
19. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017) 3
20. Liu, B., Shu, X., Wu, X.: Demoir\'eing of camera-captured screen images using deep convolutional neural network. arXiv preprint arXiv:1804.03809 (2018) 1, 3
21. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 85–100 (2018) 4
22. Liu, L., Liu, J., Yuan, S., Slabaugh, G., Leonardis, A., Zhou, W., Tian, Q.: Wavelet-based dual-branch network for image demoiréing. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 86–102. Springer (2020) 1, 2, 3, 6, 10, 11, 12
23. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 11
24. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-resolution residual networks for semantic segmentation in street scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4151–4160 (2017) 4
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 3, 4
26. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016) 8
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 11
28. Song, Y., Yang, C., Lin, Z., Liu, X., Huang, Q., Li, H., Kuo, C.C.J.: Contextual-based image inpainting: Infer, match, and translate. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018) 4
29. Sun, Y., Yu, Y., Wang, W.: Moiré photo restoration using multiresolution convolutional neural networks. IEEE Transactions on Image Processing **27**(8), 4160–4172 (2018) 1, 2, 3, 6, 11, 12, 14
30. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. arXiv preprint arXiv:2109.07161 (2021) 4
31. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8174–8182 (2018) 4
32. Vedaldi, A., Fulkerson, B.: Vlfeat: An open and portable library of computer vision algorithms. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 1469–1472 (2010) 5
33. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3349–3364 (2020) 2, 4, 7, 10

34. Wang, Y., Chen, Y.C., Tao, X., Jia, J.: Vcnet: A robust approach to blind image inpainting. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. pp. 752–768. Springer (2020) 4
35. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 11
36. Xie, C., Liu, S., Li, C., Cheng, M.M., Zuo, W., Liu, X., Wen, S., Ding, E.: Image inpainting with learnable bidirectional attention maps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8858–8867 (2019) 4
37. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6721–6729 (2017) 4
38. Yeh, R., Chen, C., Lim, T.Y., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with perceptual and contextual losses. arXiv preprint arXiv:1607.07539 **2**(3) (2016) 4
39. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015) 8
40. Yuan, S., Timofte, R., Slabaugh, G., Leonardis, A., Zheng, B., Ye, X., Tian, X., Chen, Y., Cheng, X., Fu, Z., et al.: Aim 2019 challenge on image demoireing: Methods and results. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 3534–3545. IEEE (2019) 1, 2, 3, 6, 11
41. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14821–14831 (2021) 3
42. Zhang, H., Dai, Y., Li, H., Koniusz, P.: Deep stacked hierarchical multi-patch network for image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5978–5986 (2019) 3
43. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE transactions on image processing **26**(7), 3142–3155 (2017) 3
44. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 11
45. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2472–2481 (2018) 3, 8
46. Zheng, B., Yuan, S., Slabaugh, G., Leonardis, A.: Image demoireing with learnable bandpass filters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3636–3645 (2020) 1, 2, 3, 6, 7, 10, 11, 12
47. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018) 4