Enhanced Coarse-to-Fine Network for Image Restoration from Under-Display Cameras

Yurui Zhu ⋆, Xi Wang, Xueyang Fu †, and Xiaowei Hu †

University of Science and Technology of China, Hefei, China
Shanghai AI Laboratory
{zyr, wangxxi}@mail.ustc.edu.cn, xyfu@ustc.edu.cn, huxiaowei@pjlab.org.cn

Abstract. New sensors and imaging systems are the indispensable foundation for mobile intelligent photography and imaging. Nowadays, under-display cameras (UDCs) demonstrate practical applicability in smartphones, laptops, tablets, and other scenarios. However, the images captured by UDCs suffer from complex image degradation issues, such as flare, haze, blur, and noise. To solve the above issues, we present an Enhanced Coarse-to-Fine Network (ECFNet) to effectively restore the UDC images, which takes the multi-scale images as the input and gradually generates multi-scale results from coarse to fine. We design two enhanced core components, *i.e.*, Enhanced Residual Dense Block (ERDB) and multi-scale Cross-Gating Fusion Module (CGFM), in the ECFNet, and we further introduce progressive training and model ensemble strategies to enhance the results. Experimental results show superior performance against the existing state-of-the-art methods both qualitatively and visually, and our ECFNet achieves the best performance in terms of all the evaluation metrics in the MIPI-challenge 2022 Under-display Camera Image Restoration track. Our source codes are available at this repository.

Keywords: Under-display cameras, deep neural networks, and image restoration.

1 Introduction

Under-Display Cameras (UDC) is an emerging imaging system technology, which has been applied to various scenarios, such as mobile phones, laptops, and video-conferencing with UDC TV. This novel imaging system can achieve the demands of full-screen devices with larger screen-to-body ratios, which is favored by many consumers and manufacturers. However, the display panel in front of the camera inevitably brings incoming light loss and diffraction, which in turn causes unwanted noise, blurring, fogging, and vertigo effects. The UDCs degradation process in the high dynamic scenes [7] is formulated as

$$\hat{y} \;=\; \varphi\left[C(x * k + n)\right] \;, \tag{1}$$

⋆ : This work was done during his internship at Shanghai Al Laboratory.
† : Corresponding authors.

where x, $\hat{y}$ are the original images and the UDC degradation images; $\varphi(\cdot)$ represents the non-linear tone mapping operation to reproduce the appearance of high dynamic range images on a display with more constricted dynamic range; $C(\cdot)$ is the truncation function, which limits the range of values for imaging; k refers to the Point Spread Function (PSF), which mainly brings diffraction artifacts; $*$ indicates the 2D convolution operation; n denotes the noise information from the camera device.

To remove undesired image artifacts, lots of works based on deep neural networks (DNNs) have been proposed for UDC image restoration. For example, MSUNET [45] attempts to analyze the UDC degradation processes based on two common displays, *e.g.*, the Transparent OLED (T-OLED) and the phone Pentile OLED (P-OLED). However, they just employ a U-net-based structure to explore the mapping relationship between the UDC distorted images and clean images. Feng *et al.* [7] integrate PSF prior and degradation knowledge to construct the dynamic network [14] to recover the desired high-quality images. Based on the optical properties of OLEDs, BNUDC [17] develops a two-branch structure network to achieve independent high- and low-frequency component reconstruction. Although the existing methods show remarkable performance on the restoration results in terms of multiple evaluation metrics, the local high-frequency artifacts (*e.g.*, flare caused by strong light) are still a challenging task, which largely degrades the visual results.

As mentioned before, images captured by UDC usually introduce many degradation types (*e.g.*, blurry and diffraction). There are already a large number of techniques for image restoration tasks. For example, attention and adaptive gating mechanisms [3] [27] [2] [40] [8] [43] [23] [35] [6] [19] have been widely used in low-level vision tasks, which helps the restoration network to pay extra attention to local regions. NAFNet [4] and [20] explores the applications and variants of nonlinear activation functions in their networks. Although these methods are not specifically designed for UDC restoration, they also provide potential inspiration for designing more effective UDC restoration frameworks.

In this paper, inspired by [6], we design an effective Enhanced Coarse-to-Fine Network (ECFNet) to restore images from UDCs. To be specific, we maintain the multi-input encoder and multi-output decoder UNet-based framework, which gradually restores latent clean results in different scales in a coarse-to-fine manner. Moreover, we further explore several enhanced strategies in our ECFNet to better restore the images captured by UDC. Firstly, considering the larger resolution size of UDC images, we adopt the images with multiple resolutions as the inputs of the network and output the restored image with the different resolutions. Secondly, we present an enhanced Residual Dense Block (ERDB) as the basic block of our backbone and explore the effect of different non-linear activation functions. Moreover, the multi-scale Cross-Gating Fusion Module (CGFM) is devised to merge cross-scale features to transfer the information from the multi-input encoder into the multi-output decoder with the cross-gating mechanism. Additionally, we equip with progressive training and model ensemble

strategies to further improve the restoration performance. Finally, we conduct various experiments to show the effectiveness of our method.

The main contributions of our paper are summarized as follows:

- We present an effective Enhanced Coarse-to-Fine Network (ECFNet) to restore images captured by Under-Display Cameras (UDCs) with various types of degradation.
- We devise two core enhanced modules, *i.e.*, Enhanced Residual Dense Block (ERDB) and multi-scale Cross-Gating Fusion Module (CGFM), to enhance the network by fully exploring the effects of the non-linear activation functions and the feature interaction strategies among different scales.
- Experimental results demonstrate that our method significantly outperforms other UDC image restoration solutions. In the MIPI-challenge 2022 Under-display Camera Image Restoration track, our ECFNet achieves first place in terms of all the evaluation scores ($PSNR$, $SSIM$, and $LPIPS$) and outperforms the others by a large margin.

2 Related Work

UDCs Restoration

For the development of full-screen smartphones, under-display cameras (UDCs) are a crucial technology. Due to the light loss and diffraction introduced by the front display panel, the image captured by the UDC usually suffers some degree of degradation. The UDC system has been analyzed in several previous works. An Edge Diffusion Function model for transparent OLEDs is developed by Kwon *et al.* [18]. Based on diffraction theory, Qin *et al.* [25] propose an accurate method for calculating the perspective image and demonstrate that image blur can be suppressed by modifying the pixel structure of transparent OLED displays. Qin *et al.* [26] and Tang *et al.* [29] describe and analyze the diffraction effects of the UDC system. However, these methods do not recover the already existing blurred UDC images. [44, 45] are the first work to perform UDC image restoration on publicly released datasets, by modeling the degraded images captured by the UDC, Zhou *et al.* [45] design an MCI to capture paired images, and propose a data synthesis pipeline for generating UDC images to be generated from natural images. Then based on these synthetic data, they use a variation of Unet to achieve UDC image restoration. However, they only solve the single degradation of images and are difficult to apply to UDC images with real degradation. DAGF [28] proposes a Deep Atrous Guided Filter for image restoration in UDC systems. Panikkasseril *et al.* [24] propose two different networks to recover blurred images that are captured using two types of UDC techniques. Yang *et al.* [33] use residual dense networks to achieve UDC image restoration. Kwon *et al.* [18] propose a novel controllable image restoration framework for UDC. Feng *et al.* [7] achieve restoration of multiple PSFs by adjusting the PSF of the UDC image. BNUDC [17] proposes a two-branch DNN architecture for recovering the high-frequency and low-frequency components of an image, respectively.

Coarse-to-Fine Networks

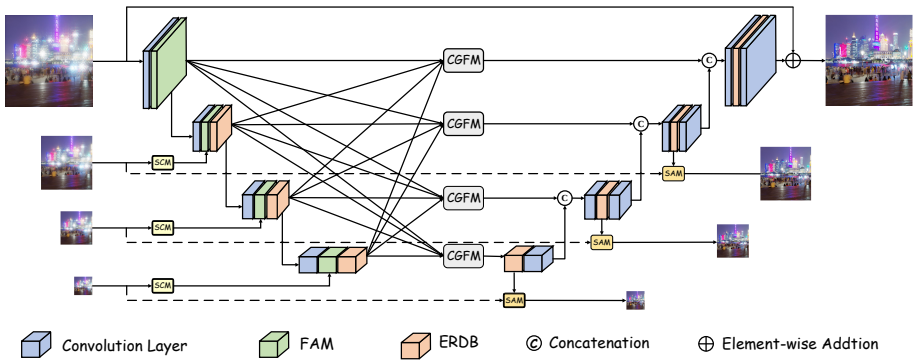

Fig. 1: The architecture of our proposed Enhanced Coarse-to-Fine Network (ECFNet) to restore the images captured by under-display cameras.

In many image restoration tasks [21] [38] [9] [30] [6], the coarse-to-fine network design has shown to be effective. For example, MIMOUet [6] rethinks the coarse-to-fine strategy and presents a UNet-like structure network to progressively recover clean images at different scales. And our network is built based on MIMOUet. However, we we have made further adjustments from the basic network module, hierarchical feature interaction manner, and network levels to accommodate to the UDCs restoration task.

3 Method

Our proposed Enhanced Coarse-to-Fine Network (ECFNet) is built with multi-input encoder and multi-output decoder UNet-based framework, which gradually generates desired high-quality results of multi-scale sizes in a coarse to fine manner. The specific architecture of our ECFNet is present in Figure 1, which is based on the MIMO-UNet [6]. Different from MIMO-UNet, we expand the scales of the network and devise the enhanced basic block to improve the network capability. Furthermore, we build the Cross-Gating Fusion Module (CGFM) to transfer the information flow among the encoders and decoders of cross scales. The specific components of our ECFNet will be thoroughly described in the following subsections.

3.1 Enhanced Encoder

The enhanced Encoder consists of four sub-networks, which take the different images with different resolutions as input. Different from MIMO-UNet [6], we further enhance the basic blocks to enhance the capacity of our model. Figure 2 depicts our proposed Enhanced Residual Dense Block, which comprises three convolution layers with kernel size of 3×3. Additionally, there are several

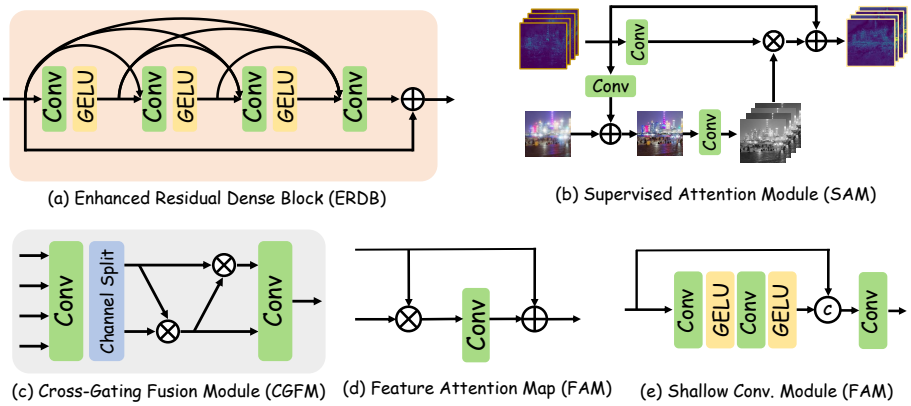

Fig. 2: The structures of sub-modules in the main architecture.

skip connections across the various levels to construct a dense network structure. Such residual dense connection manner has been proven to be effective in many computer vision tasks [13] [41] [42], which could extract richer contextual information and transfer the information flow from shallow to deep layers.

To build more effective residual dense block, we further exploit the effect non-linear activation functions. Inspired by [4], it seems that GELU [12] has more effective than other activation functions, *e.g.*, ReLU [1], PReLU [10]. Refer to [12], GELU could be approximately written as

$$GELU(x) \approx 0.5x\left(1 + tanh\left[\sqrt{2/\pi}\left(x + 0.044715x^3\right)\right]\right) \tag{2}$$

Specificlly, we replace the plain ReLU with GELU in our Enhanced Residual Dense block (ERDB), which brings obvious performance gain. Hence, we utilize the ERDB as basic block of our enhanced encoder. At each scale of the enhanced encoder, there are six ERDB in total.

3.2 Enhanced Decoder

Similar to the enhanced encoder, Enhanced Decoder (ED) also has sub-networks and takes ERDB as the basic block. Moreover, ED generate four high-quality results with different sizes and we also impose the corresponding intermediate constraints.

Besides, we further introduce the Supervised attention module (SAM) [36] between each two scales of ED, enabling for a considerable performance improvement. The specific structure of SAM is shown in Figure 2 (b).

3.3 Cross-Gating Fusion Module

For many UNet-based image restoration methods [34] [33], they usually propagate the feature information flow of the corresponding size in the encoder to the

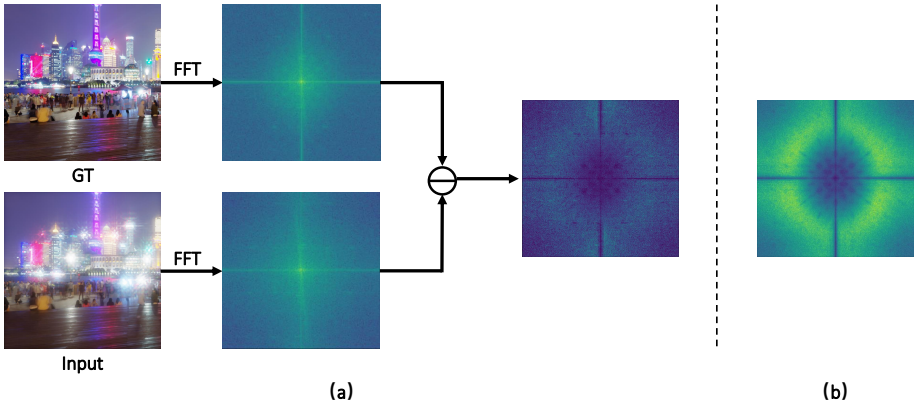

Fig. 3: Analysis of frequency domain. (a) Amplitude difference acquisition pipeline of one pair image. (b) Average amplitude difference map on the evaluation dataset.

decoder of the corresponding size to increase the diversity of features to further improve the performance of the model. Furthermore, some methods [37] [6] [15] cascade the different scale features to allow the information flow between the encoder and the decoder in a top-to-bottom and bottom-to-top manner.

In this paper, not only we adopt the dense concatenation cross features with different scales, but also we introduce the gating mechanisms to further enhance the capacity of our model. As shown in Figure 2, we propose multi-scale Cross-Gating Fusion Module (CGFM) enhance the information flow between the encoder and the decoder. CGFM first collects the features at four scales through the concatenation operation, then adjusts the number of feature channels through 1×1 convolutional layer. Then the features are evenly divided into two parts and performs gating operation [34] on the two parts sequencely. Finally, enhanced feature through the concatenation operation could be obtained. Therefore, CGFM at the smallest scale could be expressed as

$$\begin{aligned} F &= Conv_{1\times1}(Concat([(EEB_1^{out})^{\downarrow},,(EEB_2^{out})^{\downarrow},(EEB_3^{out})^{\downarrow},EEB_4^{out}])),\\ F_1, F_2 &= Split(F),\\ F_1 &= F_2 \otimes F_1,\\ F_2 &= F_1 \otimes F_2,\\ F_e &= Concat([F_1, F_2]), \end{aligned} \tag{3}$$

where $EEB_i^{out}, i = 1, 2, 3, 4$ denotes the output of the n^{th} scale enhanced encoder blocks; $Concat(\dots)$ denotes the concatenation operation along the channel dimension; $Conv_{1\times1}(\dots)$ denotes the convolution operation with kernel of 1×1;

Table 1: Quantitative comparisons of methods on the official testing datasets of the MIPI-challenge 2022 Under-display Camera Image Restoration track. The best and the second results are boldfaced and underlined, respectively. Note that ECFNet-s3 denotes that the model is obtained with the multiple training strategies.

MIPI 2022 UDC Challenge Track	Metrics		
	PSNR ↑	SSIM ↑	LPIPS ↓
ECFNet-s3 (Ours)	**48.4776**	**0.9934**	**0.0093**
2nd	<u>47.7821</u>	0.9913	0.0122
3rd	46.9227	<u>0.9929</u>	<u>0.0098</u>
4th	46.1177	0.9892	0.0159
5th	45.8722	0.9920	0.0109
6th	45.8227	0.9917	0.0106
7th	45.5613	0.9912	0.0129
8th	44.0504	0.9908	0.0120
9th	43.4514	0.9897	0.0125
10th	43.4360	0.9899	0.0133
11th	43.1300	0.9872	0.0144
12th	42.9475	0.9892	0.0150
13th	42.0392	0.9873	0.0155
14th	39.5205	0.9820	0.0216
15th	37.4569	0.9973	0.0370
16th	36.6961	0.9783	0.0326
17th	35.7726	0.9719	0.0458
18th	35.5031	0.9616	0.0453
19th	32.7496	0.9591	0.0566

$\downarrow$ denotes the down-sampling operation; $Split(\dots)$ denotes the split operation along the channel dimension; $\otimes$ denotes the element-wise multiplication.

3.4 Loss Functions

Because ECFNet has multi-scale outputs, we naturally combine the Charbonnier loss [31] and the multi-scale content losses [22] to optimize our network, which is written as follows:

$$\mathcal{L}_{content} = \sum_{k=1}^{K} \sqrt{\left\| I_{pre}^{k} - I_{gt}^{k} \right\|_2 + \epsilon^2}, \tag{4}$$

where K is the number of scales, which equals to 4 as default; I_{pre}^{k} is the predicted result of scale k; I_{gt}^{k} is the ground truth of scale k; ϵ is set to 0.0001 as default.

Besides the content loss, we also exploit the frequency domain information to provide auxiliary loss for our network. As shown in Figure 6, we first provide the visualization of amplitude difference map in frequency domain, which indicates

Table 2: Quantitative comparisons of methods on the official evaluation and testing datasets of TOLED [5]. Note that we reduce the size of the original model similar to BNUDC [17] for fair comparisons. The average inference time (IT) is obtained with TOLED dataset on the NVIDIA 1080Ti GPU device.

TOLED	PARAM(M: 10^6)	IT(seconds)	TEST SET			VAL. SET		
			PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
MSUNET [45]	8.9	0.33	37.40	0.976	0.109	38.25	0.977	0.117
IPIUer [44]	24.7	0.43	38.18	0.979	0.113	39.03	0.981	0.121
BAIDU [44]	20.0	0.18	38.23	0.980	0.102	39.06	0.981	0.111
DAGF [28]	1.1	1.30	37.27	0.973	0.141	37.46	0.943	0.090
BNUDC [17]	4.6	0.36	38.22	0.979	0.099	39.09	0.981	0.107
ECFNet-small (Ours)	4.1	0.04	37.79	0.953	0.061	38.49	0.955	0.060

Table 3: Ablation study of Basic Blocks and Modules.

Models	Experiment Setting			#Param (M: 10^6)	Metrics	
					PSNR ↑	SSIM ↑
Model-1	Basic Block	Resdiual Block	w LeakeyReLU	16.82 M	34.92	0.976
Model-2			w GELU	16.82 M	35.09	0.981
Model-3		Dense Residual Block	w ReLU	16.85 M	43.33	0.992
Model-4			w LeakeyReLU	16.85 M	44.33	0.993
Model-5			w PReLU	16.85 M	42.98	0.992
Default			w GELU	16.85 M	45.08	0.994
Model-6	Modules	w./o SAM		16.65 M	44.83	0.993
Default		w. SAM		16.85 M	45.08	0.994

the various degradations brought by UDCs mainly affect the information in the mid- and high-frequency regions. Such information loss patterns are not only obvious on a single pair of images, but also on the entire evaluation dataset. Hence, we further apply multi-scale frequency loss to optimize our network, which is defined as follows:

$$\mathcal{L}_{frequency} = \sum_{k=1}^{K} \left\| \mathcal{F}(I_{pre}^{k}) - \mathcal{F}(I_{gt}^{k}) \right\|_1, \tag{5}$$

where $\mathcal{F}(\cdot)$ indicates the Fast Fourier Transform (FFT). Finally, the total loss could be defined as

$$\mathcal{L}_{total} = \mathcal{L}_{content} + \lambda \mathcal{L}_{frequency}, \tag{6}$$

where λ denotes the balanced weight and we empirically set λ to 0.5 as default.

4 Experiments

4.1 Implementation Details

We implement our proposed UDC image restoration network via the PyTorch 1.8 platform. Adam optimizer [16] with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$ is

Table 4: Ablation study of the losses.

Losses			Metrics	
L1 Loss	Charbonnier L1 Loss	Frequency Loss	PSNR ↑	SSIM ↑
✓			42.26	0.989
	✓		42.50	0.990
✓		✓	44.83	0.993
	✓	✓	45.08	0.994

Table 5: Ablation study of the connection manner between Enhanced Encoder (EE) and Enhanced Decoder (ED).

Connection Manner Between EE and ED	Metrics	
	PSNR ↑	SSIM ↑
Simple Concatentation	44.00	0.991
Single Gating Fusion	44.87	0.993
Cross-Gating Fusion Module (CGFM) (default)	45.08	0.994

adopted to optimize our network. Additionally, motivated by [34], we introduce the progressive training strategy and the specific training phase of our network could be divided into three stages:

(1) We adopt the Adam optimizer with a batch size of three and the patch size of 256×256. The initial learning rate is 2×10^{-4} and is adjusted with the Cosine Annealing scheme, including 1000 epochs in total. The first stage performs on the NVIDIA 1080Ti GPU. We obtain the best model at this stage as the initialization of the second stage.

(2) We adopt the Adam optimizer with a batch size of one and the patch size of 512 × 512. The initial learning rate is 1×10^{-5} and is adjusted with the Cosine Annealing scheme, including 300 epochs in total. The second stage performs on the NVIDIA 1080Ti GPU. We obtain the best model at this stage as the initialization of the next stage.

(3) We adopt the Adam optimizer with a batch size of two and the patch size of 800 × 800. The initial learning rate is 8×10^{-6} and is adjusted with the Cosine Annealing scheme, including 150 epochs in total. The third stage performs on the NVIDIA A100 GPU. In this stage, we additionally expand the training dataset with the testing set of SYNTH Dataset [7]. Finlly, We adopt the model ensemble strategy, which averages the parameters of multiple models trained with different training iterations.

Therefore, in order to better distinguish the results of the models, we record the results of the three stages as ECFNet-s1, ECFNet-s2 and ECFNet-s3.

In the testing phase, we adopt the model after fine-tuning to achieve the best performance. Moreover, we utilize a model-ensemble strategy to obtain better results. 11G GPU memory is enough to infer our model, and we use one NVIDIA 1080Ti GPU with 11G memory for testing.

Table 6: Ablation study of the numbers of network inputs and ouputs.

Numbers of Network Inputs and Outputs	Metrics	
	PSNR ↑	SSIM ↑
Single Input, Single Output	44.49	0.992
Three Inputs, Three Outputs	44.68	0.993
Four Inputs, Four Outputs (default)	45.08	0.994

Table 7: Ablation study of training strategies.

Results with Different Strategy	Metrics	
	PSNR ↑	SSIM ↑
ECFNet-s1	45.08	0.994
ECFNet-s2 + w./o Externel Data	46.23	0.995
ECFNet-s2 + w. Externel Data	47.80	0.996
ECFNet-s3 + w. Externel Data	49.04	0.995
ECFNet-s3 + w. Externel Data + Model Ensemble (default)	49.13	0.996

4.2 Dataset

We conduct the experiments strictly following the setting of the MIPI-challenge 2022 Under-display Camera Image Restoration track. There are 2016 pairs of $800 \times 800 \times 3$ images in the training split. Image values are produced in ".npy" format and range from [0, 500]. There are 40 image pairings in the validation set and 40 image pairs in the testing dataset, respectively. Note that the corresponding ground truths of testing dataset are not publicly available. Besides the above dataset at the traning phase, we also exploit an additional 320 pairs of test datasets in dataset SYNTH [7] to augment the training dataset. However, our network never requires the Point Spread Function (PSF) as the externel prior to predict the final results, which is more flexible than DISCNet [7]. Only for the challenge dataset, we adopt the additional dataset to improve the restoration performance of our model.

Moreover, in order to validate the effectiveness of our method, we also employ the TOLED dataset [5] to training our network. TOLED dataset totally consists of 300 image pairs, which has been divided into 240 image pairs for training, 30 image pairs for evaluation, and 30 image pairs for testing. Note that the models trained on these two different datasets are different, and then applied to the corresponding test datasets for testing.

4.3 Evaluation Metrics

Similar to previous methods [7] [17], we employ three reference-based metrics to verify the effectiveness of our method: Peak Signal-to-Noise Ratio (PSNR), the structural similarity (SSIM) [32], and Learned Perceptual Image Patch Similarity (LPIPS) [39]. For the PSNR and SSIM metrics, higher is better. For the LPIPS metric, lower means better.

Fig. 4: Visual comparison results of UDC Image Restoration on the evaluation dataset of MIPI-challenge 2022 track. Note that brighter means bigger error.

4.4 Comparations

In Table 1, we report the comparison results among different solutions on the challenge track. Obviously, our method performs best performance in terms of all the evaluation metrics, even 0.7 dB higher than the second-place method. Besides, in table 2 we achieve the comparable performance on the TOLED evaluation and testing dataset. For fair comparison, note that we reduce the size of the original model similar to BNUDC [17]. And our method is significantly seventeen times faster than BNUDC on the TOLED dataset. In addition, in Figure 4,5,6, to demonstrate that the images generated by our network have better visual quality, we show the residual map of other methods and our on the UDC dataset to increase the visual differentiation. Note that the residual map is the difference between the estimated results and the ground truth. It is clear that our method achieves better visual results, with better recovery for edge flare and less difference between the generated images and the ground truth.

Fig. 5: Visual comparison results of UDC Image Restoration on the testing dataset of MIPI-challenge 2022 track.

4.5 Ablation Study

We conduct extensive experiments to verify the effect of each components of our method, *e.g.*, modules, strategies and losses. Note that in the ablation study we only use the first stage training manner (refer to ECFNet-s1) for convenience.

Effects of Basic Blocks and Modules AS reported in Table 3, we compare the performance of models with different non-linear activation functions, including ReLU, LeakeyReLU, PReLU, and GELU. Obviously, using GELU performs better than other activation functions both embedding in the standard residual block [11] and standard dense residual block [13]. For example, compared to using LeakeyReLU, using GELU brings 0.75db performance gains.

In Table 3, we further conduct experiments to verify the effectiveness of the SAM. Using SAM [36] aims at bridging the relation between the different scales, which facilitates achieving 0.25 dB performance gain.

Effects of the Coarse-to-Fine Strategy Our restoration framework is built on the coarse-to-fine strategy to gradually recover different scales latent clean results. We further report the related results of the model with different numbers of inputs and outputs in Table 6, which could validate the effectiveness of the

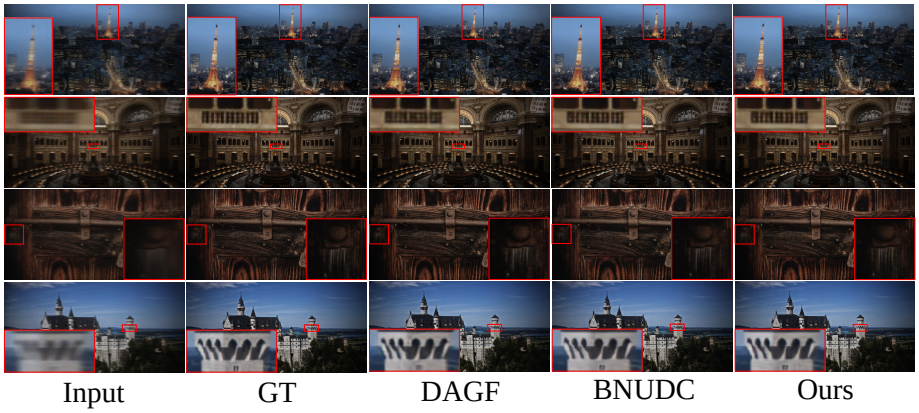

Fig. 6: Visual comparison results of UDC Image Restoration on the TOLED dataset.

coarse-to-fine strategy. Obviously, using four inputs and four outputs achieve the best results. Compared with baseline, using multiple outputs to optimize our model brings about 0.59 dB performance improvement.

Effects of the Loss Functions Except for the content losses, we additionally introduce the $\mathcal{L}_{frequency}$ to optimize the network from frequency domain. The network performance with different losses is reported in Table 4. Furthermore, we provide the visualization comparisons when models are optimized by different loss combinations in Figure 7. Combining Charbonnier L1 Loss and Frequency Loss brings the best model performance.

Effects of the Training Strategies Following Restormer [34], we additionally adopt the progressive training strategy to enhance the model performance. As shown in Table 7, the models trained at different stages are marked as ECFNet-s1, ECFNet-s2, and ECFNet-s3. Experiments show that training with progressively larger patches often leads to better generalization performance gains

Model ensemble strategy, whose results are obtained by linearly combine several model parameters with different training iterations. This brings around a 0.09dB+ (PSNR) increase on the evaluation dataset with help of the model ensemble strategy.

Inference Time. The average inference time per image of our standard ECFNet is 0.2665s, which is obtained with the image resolution of 800 $\times$ 800 on the NVIDIA 1080Ti GPU device.

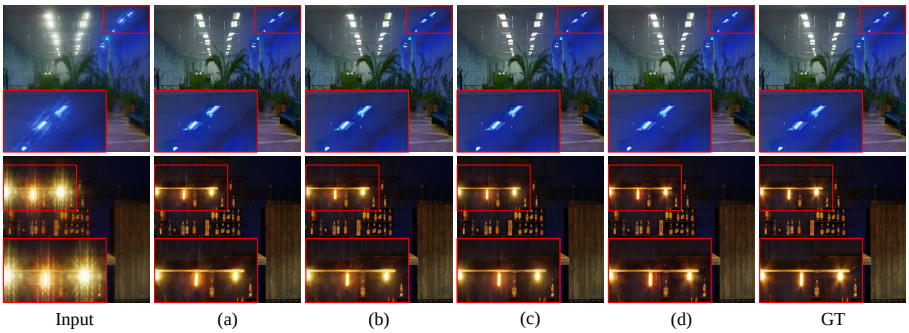

Fig. 7: Model results with different loss combinations. (a) w. L1 Loss; (b) w. Charbonnier L1 Loss; (c) w. L1 Loss and FFT loss; (d) w. Charbonnier L1 Loss and FFT loss (default).

5 Conclusions

In this paper, we design an effective network for restoring images from under-display cameras, named Enhanced Coarse-to-Fine Network (ECFNet). It inherits the previous classic coarse-to-fine network framework and further formulates two enhanced core modules, *i.e.*,, Enhanced Residual Dense Block (ERDB) and multi-scale Cross-Gating Fusion Module (CGFM). Besides, we introduce the progressive training and model ensemble strategy to further improve our model performance. Finally, ECFNet achieves the best performance in terms of all the evaluation metrics in the MIPI-challenge 2022 UDC Image Restoration track.

Acknowledgement This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61901433 and in part by the USTC Research Funds of the Double First-Class Initiative under Grant YD2100002003. This work is partially supported by the Shanghai Committee of Science and Technology (Grant No.21DZ1100100).

References

1. Agarap, A.F.: Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375 (2018) 5
2. Anwar, S., Barnes, N.: Real image denoising with feature attention. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3155–3164 (2019) 2
3. Chen, D., He, M., Fan, Q., Liao, J., Zhang, L., Hou, D., Yuan, L., Hua, G.: Gated context aggregation network for image dehazing and deraining. In: 2019 IEEE winter conference on applications of computer vision (WACV). pp. 1375–1383. IEEE (2019) 2
4. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. European Conference on Computer Vision (2022) 2, 5
5. Cheng, C.J., Huang, T.C., Lin, W.T., Hsieh, C.C., Chen, P.Y., Lu, P., Lin, H.Y.: P-79: Evaluation of diffraction induced background image quality degradation through transparent oled display. In: SID Symposium Digest of Technical Papers. vol. 50, pp. 1533–1536. Wiley Online Library (2019) 8, 10
6. Cho, S.J., Ji, S.W., Hong, J.P., Jung, S.W., Ko, S.J.: Rethinking coarse-to-fine approach in single image deblurring. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4641–4650 (2021) 2, 4, 6
7. Feng, R., Li, C., Chen, H., Li, S., Loy, C.C., Gu, J.: Removing diffraction image artifacts in under-display camera via dynamic skip connection network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 662–671 (2021) 1, 2, 3, 9, 10
8. Fu, X., Qi, Q., Zha, Z.J., Zhu, Y., Ding, X.: Rain streak removal via dual graph convolutional network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1352–1360 (2021) 2
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) 4
10. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015) 5
11. 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) 12
12. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016) 5
13. 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) 5, 12
14. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. Advances in neural information processing systems **29** (2016) 2
15. Kim, S.W., Kook, H.K., Sun, J.Y., Kang, M.C., Ko, S.J.: Parallel feature pyramid network for object detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 234–250 (2018) 6
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 8
17. Koh, J., Lee, J., Yoon, S.: Bnudc: A two-branched deep neural network for restoring images from under-display cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1950–1959 (2022) 2, 3, 8, 10, 11

18. Kwon, H.J., Yang, C.M., Kim, M.C., Kim, C.W., Ahn, J.Y., Kim, P.R.: Modeling of luminance transition curve of transparent plastics on transparent oled displays. Electronic Imaging **2016**(20), 1–4 (2016) 3
19. Liu, D., Wen, B., Fan, Y., Loy, C.C., Huang, T.S.: Non-local recurrent network for image restoration. Advances in neural information processing systems **31** (2018) 2
20. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022) 2
21. Ma, Y., Liu, X., Bai, S., Wang, L., He, D., Liu, A.: Coarse-to-fine image inpainting via region-wise convolutions and non-local correlation. In: IJCAI. pp. 3123–3129 (2019) 4
22. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3883–3891 (2017) 7
23. Pan, X., Zhan, X., Dai, B., Lin, D., Loy, C.C., Luo, P.: Exploiting deep generative prior for versatile image restoration and manipulation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 2
24. Panikkasseril Sethumadhavan, H., Puthussery, D., Kuriakose, M., Charangatt Victor, J.: Transform domain pyramidal dilated convolution networks for restoration of under display camera images. In: European Conference on Computer Vision. pp. 364–378. Springer (2020) 3
25. Qin, Z., Tsai, Y.H., Yeh, Y.W., Huang, Y.P., Shieh, H.P.D.: See-through image blurring of transparent organic light-emitting diodes display: calculation method based on diffraction and analysis of pixel structures. Journal of Display Technology **12**(11), 1242–1249 (2016) 3
26. Qin, Z., Xie, J., Lin, F.C., Huang, Y.P., Shieh, H.P.D.: Evaluation of a transparent display's pixel structure regarding subjective quality of diffracted see-through images. IEEE Photonics Journal **9**(4), 1–14 (2017) 3
27. Ren, W., Ma, L., Zhang, J., Pan, J., Cao, X., Liu, W., Yang, M.H.: Gated fusion network for single image dehazing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3253–3261 (2018) 2
28. Sundar, V., Hegde, S., Kothandaraman, D., Mitra, K.: Deep atrous guided filter for image restoration in under display cameras. In: European Conference on Computer Vision. pp. 379–397. Springer (2020) 3, 8
29. Tang, Q., Jiang, H., Mei, X., Hou, S., Liu, G., Li, Z.: 28-2: Study of the image blur through ffs lcd panel caused by diffraction for camera under panel. In: SID Symposium Digest of Technical Papers. vol. 51, pp. 406–409. Wiley Online Library (2020) 3
30. Wang, L., Li, Y., Wang, S.: Deepdeblur: fast one-step blurry face images restoration. arXiv preprint arXiv:1711.09515 (2017) 4
31. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019) 7
32. 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) 10
33. Yang, Q., Liu, Y., Tang, J., Ku, T.: Residual and dense unet for under-display camera restoration. In: European Conference on Computer Vision. pp. 398–408. Springer (2020) 3, 5

34. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739 (2022) 5, 6, 9, 13
35. 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) 2
36. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: CVPR (2021) 5, 12
37. 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) 6
38. Zhang, K., Tao, D., Gao, X., Li, X., Li, J.: Coarse-to-fine learning for single-image super-resolution. IEEE transactions on neural networks and learning systems **28**(5), 1109–1122 (2016) 4
39. 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) 10
40. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018) 2
41. 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) 5
42. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(7), 2480–2495 (2020) 5
43. Zhao, H., Kong, X., He, J., Qiao, Y., Dong, C.: Efficient image super-resolution using pixel attention. In: European Conference on Computer Vision. pp. 56–72. Springer (2020) 2
44. Zhou, Y., Kwan, M., Tolentino, K., Emerton, N., Lim, S., Large, T., Fu, L., Pan, Z., Li, B., Yang, Q., et al.: Udc 2020 challenge on image restoration of under-display camera: Methods and results. In: European Conference on Computer Vision. pp. 337–351. Springer (2020) 3, 8
45. Zhou, Y., Ren, D., Emerton, N., Lim, S., Large, T.: Image restoration for under-display camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9179–9188 (2021) 2, 3, 8