

Point-and-Shoot All-in-Focus Photo Synthesis from Smartphone Camera Pair

Xianrui Luo, Juewen Peng, Weiyue Zhao, Ke Xian, Hao Lu, and Zhiguo Cao

Abstract—All-in-Focus (AIF) photography is expected to be a commercial selling point for modern smartphones. Standard AIF synthesis requires manual, time-consuming operations such as focal stack compositing, which is unfriendly to ordinary people. To achieve point-and-shoot AIF photography with a smartphone, we expect that an AIF photo can be generated from one shot of the scene, instead of from multiple photos captured by the same camera. Benefiting from the multi-camera module in modern smartphones, we introduce a new task of AIF synthesis from main (wide) and ultra-wide cameras. The goal is to recover sharp details from defocused regions in the main-camera photo with the help of the ultra-wide-camera one. The camera setting poses new challenges such as parallax-induced occlusions and inconsistent color between cameras. To overcome the challenges, we introduce a predict-and-refine network to mitigate occlusions and propose dynamic frequency-domain alignment for color correction. To enable effective training and evaluation, we also build an AIF dataset with 2686 unique scenes. Each scene includes two photos captured by the main camera, one photo captured by the ultra-wide camera, and a synthesized AIF photo. Results show that our solution, termed EasyAIF, can produce high-quality AIF photos and outperforms strong baselines quantitatively and qualitatively. For the first time, we demonstrate point-and-shoot AIF photo synthesis successfully from main and ultra-wide cameras.

Index Terms—All-in-Focus synthesis, main/ultra-wide camera, occlusion-aware networks

I. INTRODUCTION

ALL-IN-FOCUS (AIF) synthesis is commonly used in photography to keep everything sharp in a scene. To achieve AIF using a regular lens, one can decrease either the aperture size or the focal length. However, in smartphone photography, the focal length or the aperture size of the camera is fixed. Therefore, when the distance between the main camera and the foreground object is close, the lens will inevitably produce shallow depth-of-field (DOF) where either the foreground or background region is out-of-focus.

In AIF synthesis, current approaches fuse the focal stack [1], [2], [3], *i.e.*, a set of images shot by the same camera at different focal distances, to generate an AIF photo. However,

capturing a focal stack is time-consuming and requires repetitive manual refocusing. Can AIF synthesis be made simpler? We show that an additional camera can largely simplify AIF synthesis and extend the limited DOF of the main camera.

Specifically, telephoto/wide lens or wide/ultra-wide lens pair on smartphones constitutes the regular main/sub camera combination. Here we focus on the setting of wide/ultra-wide lens. As shown in Fig. 2, the wide-angle main lens produces high-quality details with harmonious colors, but it has a rather shallow DOF. An ultra-wide-angle lens has a small aperture and a short focal length, which results in a large DOF and an almost sharp image. It is thus natural to seek whether the ultra-wide-angle lens can help recover missing details in the main lens to achieve AIF synthesis. The advantages of wide/ultra-wide cameras in AIF synthesis are: 1) two cameras can work simultaneously, which is faster than adjusting focal distances with a single camera; 2) the second camera provides additional information for restoring defocused regions. When we capture the main/ultra-wide image pair, we only need to adjust the focal distance of the main lens and leave the ultra-wide lens to the smartphone defaults. The hardware characteristics of the main/ultra-wide pair are used to synthesize an all-in-focus image with good quality.

In this work, our goal is to extend the DOF of the main camera by exploiting the ultra-wide camera. However, as shown in Fig. 3, the following visual challenges need to be solved: 1) the large spatial displacement between two photos; 2) the illumination and color difference between two cameras; 3) the high-resolution input size for inference.

To process two inputs from different cameras, reference-based super-resolution methods [4], [5], [6], [7] are proposed to tackle spatial misalignment. However, they do not focus on shallow DOF scenes, which means they are not designed to deal with defocus blur. Furthermore, in dual-camera super-resolution, the distances between the lens and the objects of a scene are sufficiently large to define the objects as if they are from the same depth plane. On the other hand, our task deals with large foreground occlusions and defocus blur so that we cannot directly use dual-camera super-resolution models. Defocus deblurring methods [8], [9], [10], [11] aim to restore details from out-of-focus regions, but the lack of reference results in failure when the blurring amount is strong. Although current methods introduce dual-pixel image pairs, the image pairs are equivalent to stereo image pairs with a small baseline, which still provides no sharp reference for deblurring.

Contrary to existing methods, we consider all problems above and present EasyAIF, a feasible framework to synthesize AIF photos, featured by spatial alignment, color adjustment,

This work was supported in part by the National Natural Science Foundation of China (Grant No. U1913602) and funded by Huawei Technologies CO., LTD. (Corresponding author: Zhiguo Cao).

Xianrui Luo, Juewen Peng, Weiyue Zhao, Hao Lu, and Zhiguo Cao are with the Key Laboratory of Image Processing and Intelligent Control, Ministry of Education, and also with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: xianruiluo@hust.edu.cn; juewenpeng@hust.edu.cn; zhaoweiyue@hust.edu.cn; hlu@hust.edu.cn; zgcao@hust.edu.cn).

Ke Xian is with the S-Lab for Advanced Intelligence, Nanyang Technological University, Singapore (e-mail: ke.xian@ntu.edu.sg).

Digital Object Identifier 10.1109/TCSVT.2022.3222609



Figure 1. **All-in-Focus (AIF) photo synthesis from a smartphone camera pair.** We present a novel approach EasyAIF that operates two images captured simultaneously and respectively by a wide camera and an ultra-wide camera to repair large defocus blur (in the main lens) with clear contents (from the ultra-wide lens), enabling AIF synthesis with one click on the phone.

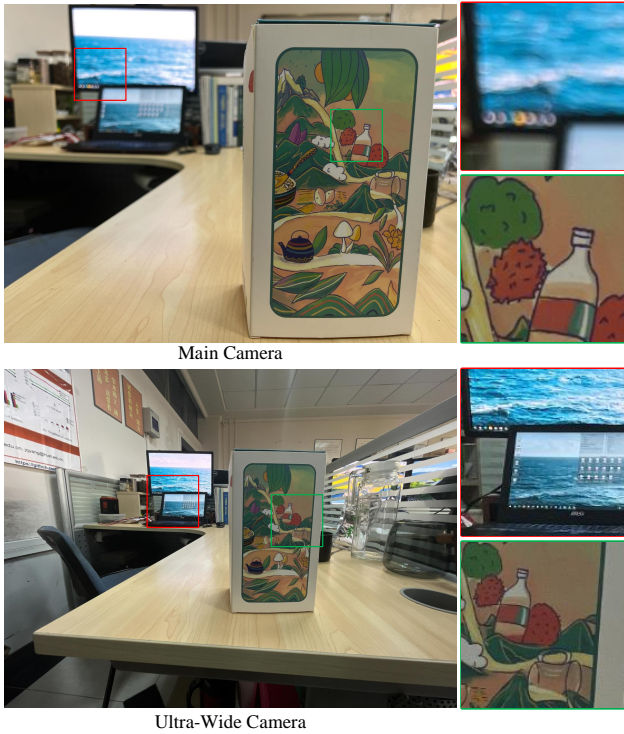


Figure 2. **Features of the smartphone main/ultra-wide camera pair.** The main lens can capture a scene with good quality, but it has a shallow depth-of-field (the red square), which results in defocus blur. Therefore, we collect the image captured by the ultra-wide lens, which provides wider depth-of-field images (the green square) compared with the main lens.

and occlusion-aware synthesis. To align spatial differences, we first apply homography-warping and flow warping. To further fix foreground occlusions, we design an occlusion-aware network with a deformable wavelet-based module aiming to generate sharp results in occluded regions.

In addition to spatial displacement, we also address the color

differences engendered by different lenses. To match the color of the ultra-wide photo to that of the main-camera photo, we propose a wavelet-based dynamic convolution network where dynamic convolution is used to predict weights conditioned on the blurred main-camera image. To better preserve information during downsampling and improve efficiency during inference, we apply wavelet transformation, which loses less information than conventional downsampling. Finally, the occlusion-aware network fuses sharp regions from the aligned images and outputs AIF results.

To train our network, we collect a dataset of 2686 scenes. We capture three photos on each scene: a background-blurred main-camera photo focused on foreground, a foreground-blurred main-camera photo focused on background, and a sharp ultra-wide sub-camera photo captured with a small aperture and a short focal length. Since the focal length and the aperture of the smartphone camera are fixed, we cannot obtain AIF ground truths for the main image directly. Therefore we resort to a fusion-based approach to synthesize a reference-based smartphone AIF dataset. In particular, multi-focus image fusion [12] is used to produce sharp ground truths by fusing the two main-camera photos.

As shown in Fig. 1, we are the first to explore the multi-camera module in smartphone AIF synthesis. The off-the-shelf ultra-wide camera provides the defocused main photo with sharp guidance. To overcome the alignment issue, we propose a viable framework to alleviate blurring artifacts. We conduct experiments to compare our solution with existing dual-camera super-resolution and defocus deblurring approaches. We observe that current baselines are not good enough for AIF synthesis. Our framework EasyAIF is tailored to AIF synthesis and outperforms other approaches qualitatively and quantitatively. It can serve as a strong baseline for AIF synthesis from main/ultra-wide camera pair.

Our main contributions include the following:

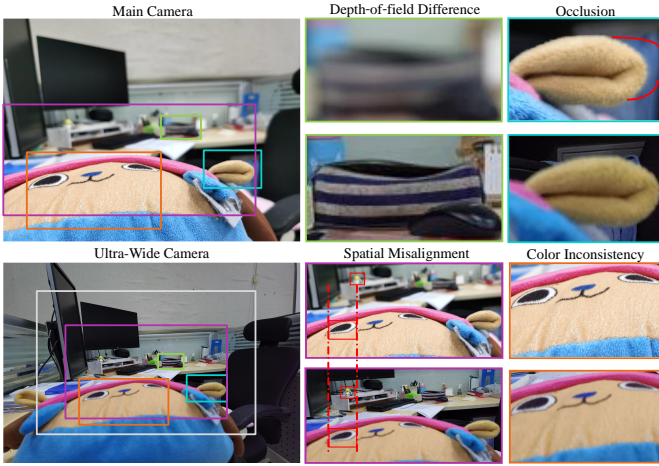


Figure 3. **Visual challenges of AIF synthesis using the main/ultra-wide image pair from a smartphone.** The wide and ultra-wide images exhibit significant visual differences in the field of view (FoV) and resolution. Although the main (wide) and ultra-wide images capture overlapped scenes (the white box), they differ in multiple aspects, including the occlusion (the red-curve area in the blue box), the spatial misalignment (the red boxes in the purple box), the difference in the depth-of-field (the green box), and color inconsistency (the orange box).

- To our knowledge, we are the first to introduce the task of AIF synthesis from the main and ultra-wide camera pair;
- A strong baseline that leverages an occlusion-aware framework which fuses the main image, the spatial-color aligned ultra-wide image, and the refined main image to produce a pleasant AIF result;
- We collect an AIF dataset with quadruplet samples where each is composed by a pair of main-camera images that respectively focus on foreground and background, an ultra-wide-camera image, and a synthetic AIF image used as ground truth.

II. RELATED WORK

A. All-In-Focus Synthesis

Current AIF synthesis approaches use a focal stack to composite the AIF image [1], [2], [3]. Typically multi-focus fusion consists of focus measure and image fusion. For sharpness measure, one can apply spatial cues [13], multi-scale decomposition [14], and sparse representation [15]. Then fusion methods such as adaptive noise-robust fusion [2] are used to selectively blend the focal stack. Apart from common solutions, the AIF image can also be reconstructed by light field synthesis [16]. Since capturing focal stack requires sweeping the focal plane, depth from focus [17], [18], [19] has been proposed to perform depth estimation along with AIF synthesis.

Recently deep learning-based image fusion is used for AIF synthesis. The focus measure and the fusion rule can be learned from a deep network [20], [21]. Deep unsupervised learning is widely used to fuse a focal stack without explicit supervision [22], [23]. In addition, depth-from-focus networks [24] adopt supervised learning to predict an AIF image and a depth map from corresponding synthetic ground truth. A joint multi-level feature extraction-based CNN [25]

is proposed to provide a natural enhancement. A simplified stationary wavelet transform using Harr filter [26] is proposed to achieve fast fusion speed.

Although focal stack is effective in generating AIF results, it is time-consuming and inefficient to manually change focal planes and capture required image sequences using the same smartphone lens. Therefore, we propose a user-friendly AIF synthesis routine to produce an AIF image from a main and ultra-wide image pair that can be captured at the same time. Traditionally the focal stack or the light field images are shot by the same camera, so the captured images do not suffer from misalignment in space and color. Compared with the common multi-focus fusion methods [25], [26], our inputs are two misaligned images from two different cameras instead of the common defocused images from the same camera.

B. Defocus Deblurring

In addition to direct AIF synthesis by image fusion, defocus deblurring is also feasible to restore a sharp image from a bokeh image [27], [28]. Traditional methods apply a two-stage technique [29], [30], [31], [32], where defocus estimation [30], [32], [33], [34], [35] is first executed on a pre-defined blur model, then the predicted defocus map helps to deblur the image by non-blind deconvolution [29], [31].

Instead of using deep networks to predict a defocus map, recent deep learning methods directly deblur the image [8], [9], [10], [11]. Therefore, a defocus deblurring dataset [8] consisting of dual-pixel images is introduced for training. Better network designs have been proposed, such as iterative adaptive [9] and kernel-sharing parallel atrous [10] convolutions. With the help of dual-pixel images, multi-task learning methods such as defocus estimation [36] and predicting dual-pixel views [37] are also proved to benefit defocus deblurring. Synthetic datasets such as dual-pixel video sequences [38] and depth images [39] have also been used as assistance for defocus deblurring. The transformer architecture is also applied to defocus deblurring [11] and achieves the state of the art. In addition, light field data [40] can be applied to synthesize defocus blur from a set of sharp images, and a network is proposed to deblur a single spatially varying defocused image.

Compared with reference-based AIF synthesis, defocus deblurring also restores a sharp image. However, it only utilizes a single image or a dual-pixel image pair instead of a sharp reference image, which is not sufficient to recover details from large blur. Compared with the restoration method from light field [40], our method requires two complimentary inputs, while the light field dataset only enables single image deblurring by synthesize one defocused image from multiple light field images. Furthermore, the light field dataset is captured by the same camera, so the light field images are aligned, which is different from our AIF dataset.

C. Dual Camera Applications

Dual camera has been used in various low-level applications such as reference-based super-resolution [4], [5], [6], [7],

reference-guided image inpainting [41], and bokeh rendering [42]. Reference-based super-resolution methods align two images from different viewpoints [4], [5], [6]. Particularly, dual-camera super-resolution [7] super-resolves the wide image with the help of the telephoto lens, which can be applied to smartphone images.

Compared with our wide/ultra-wide AIF synthesis task, dual-camera super-resolution methods are designed for two misaligned input images. However, they do not consider defocus blur or foreground occlusions. In addition, reference-guided inpainting requires a mask for inputs, and bokeh rendering does not belong to image restoration task, so there is little similarity between our task and existing ones.

III. EASYAIF FOR ALL-IN-FOCUS SYNTHESIS

As shown in Fig. 4, our framework EasyAIF consists of three components: spatial alignment, color alignment, and occlusion-aware synthesis. To solve the spatial misalignment between the main camera I_m and the ultra-wide camera I_w , we apply both homography warping and flow warping to output the aligned ultra-wide $I_{w,f}$. Homography warping focuses on a global scale and produces a coarse result, and flow warping is stable under parallax issues brought by different depth planes. Once spatial warping is done, we also align the two images I_m and $I_{w,f}$ photometrically. To process images of high resolution, we propose a wavelet-based network to preserve information during downsampling. Furthermore, we apply dynamic convolution to generate the color-aligned $I_{w,c}$ utilizing the reference blurred main image I_m . Since the dual-camera setting leads to occlusions around the edge of foreground objects, we design an occlusion-aware synthesis network, where we simultaneously find the occlusions and refine the blurred occluded area to produce I_d . Finally the fusion network fuses I_m , $I_{w,c}$, and I_d to synthesize the AIF image I_{AIF} .

A. Spatial Alignment

We implement spatial alignment by means of homography warping and flow warping. Homography warping aligns the input at the image level, which can restore image-level attributes such as the FOV. However, in our smartphone AIF synthesis task, prominent foreground/background relationships result in a larger disparity on background regions than on foreground ones. Therefore it is impossible to align the main/ultra-wide image pair I_m and I_w using only a single homography warping. To align objects from different depth planes, we apply pixel-wise warping with an optical flow field, because we need diverse offsets to fit the increasing parallax from foreground to background.

Homography warping is executed by image registration. Image registration consists of 1) keypoints detection, 2) feature matching, 3) outlier pre-filtering, and 4) pose estimation. We locate keypoints from SIFT [43], and extract HardNet [44] descriptors to compute an initial correspondences set between I_w and I_m . Then we apply a pre-trained outlier rejection network NM-Net [45] to filter unreliable correspondences in the initial

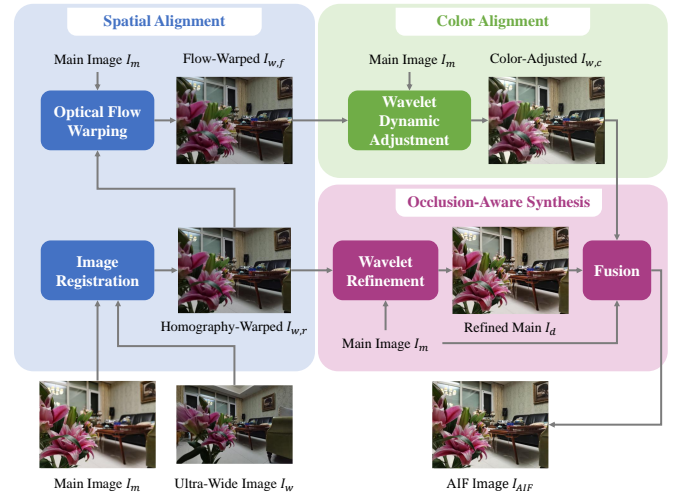


Figure 4. **Our framework EasyAIF includes three modules: spatial alignment, color alignment, and occlusion-aware synthesis.** We use homography and flow warping to align the main/ultra-wide image pair I_m and I_w at the image level and the pixel level. Then we adjust the color of the warped ultra-wide image $I_{w,f}$ with a wavelet-based dynamic network. To tackle the occlusions where spatial warping fails to solve, we propose an occlusion-aware synthesis network to refine the occluded and blurred areas. The network fuses the refined image I_d , the main image I_m , and the color-adjusted ultra-wide $I_{w,c}$ to generate the AIF result I_{AIF} .

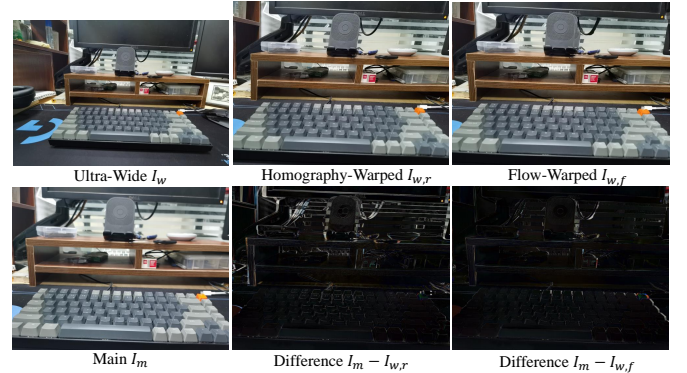


Figure 5. **The difference between a warped ultra-wide image and a main image.** I_w has a different aspect ratio from I_m .

set. Finally, I_w is warped by a single homography matrix estimated from the filtered correspondences using RANSAC [46]. Homography warping provides a coarse image-level alignment for our pipeline. The warping results $I_{w,r}$ and I_m still have unaligned regions due to large parallax. Therefore, we apply a pixel-wise warping alignment with an optical flow field. Pixel-wise warping is insensitive to varying parallax from different depth planes. We apply the robust RAFT [47] to estimate the flow map $Y_{I_m \rightarrow I_{w,r}}$ for aligning $I_{w,r}$ to I_m . As shown in Fig. 5, homography alignment provides a coarse image-level result, which is effective in dealing with FOV discrepancy, and the flow warping exhibits more adaptability in aligning objects from different depth planes such as the speaker in $I_{w,r}$.

B. Wavelet-Based Dynamic Color Alignment

In image alignment, we not only consider spatial differences, but also color discrepancies. Since two different cameras are involved, color alignment must be considered.

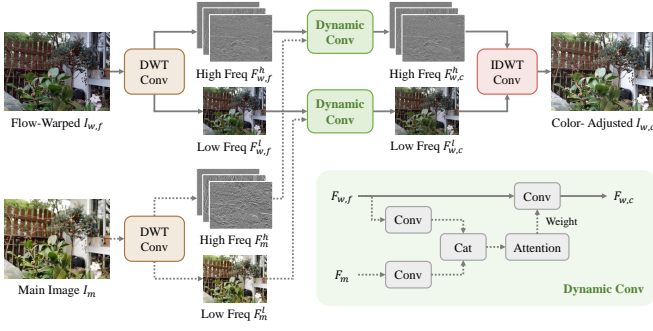


Figure 6. **WDC-Net for color alignment.** We use DWT to downsample the RGB image into high and low frequency contents F^h and F^l . To utilize the reference image, we apply dynamic convolution on high and low frequency to adjust color transform parameters based on the reference I_m . The adjusted low frequency component is reconstructed to the original resolution $I_{w,c}$ guided by high frequency details.

Specifically, smartphone AIF synthesis defines a special scenario for color alignment. In our pipeline we have a reference blurred input I_m as guidance, which is different from the previous single image enhancement [48]. Furthermore, the to-be-adjusted $I_{w,f}$ is spatially misaligned but only in occlusions, so it is not the case of style transfer [49].

To use the reference main camera I_m , we propose a wavelet-based dynamic color alignment network WDC-Net. Particularly, we integrate dynamic convolution [50] with discrete wavelet transformation (DWT) and inverse DWT. As shown in Fig. 6, DWT decomposes the RGB image into high and low frequency contents, and the resolution of output is half of the input size. It has been revealed that in image-to-image translation, the color transformation relies more on the low frequency contents [51]. Therefore, we use DWT to align the color in low frequency and refine the low frequency result with high frequency cues. Furthermore, we expect that the main image can guide color alignment, so we apply dynamic convolution such that the color-transform-convolution weight is flexible based on the input main image.

C. Occlusion-Aware Synthesis

Now that we discuss the spatial alignment performed by homography and flow warping, we still have the occlusion problem. Occlusion arisen from two perspectives is not rare. However, it is worth emphasizing in smartphone AIF synthesis. As shown in Fig. 3, when the camera is close to the foreground object, the parallaxes in each depth plane are significantly different, which brings more occlusions than other tasks such as dual-camera super-resolution.

Taking occlusions into account, when the main camera focuses on foreground objects, the ultra-wide image cannot obtain corresponding sharp results in the occluded regions. The flow warping in spatial alignment tends to produce poor flows in occluded areas, so we first predict a confidence map M_c by forward-backward consistency check [52] as

$$M_c = \|Y_{I_{w,r} \rightarrow I_m}(p) + Y_{I_m \rightarrow I_{w,r}}(p + Y_{I_{w,r} \rightarrow I_m}(p))\|_2 \leq 1, \quad (1)$$

where p is considered as a point in $I_{w,r}$, $Y_{I_{w,r} \rightarrow I_m}$ is the flow from the ultra-wide image to the main image, and $Y_{I_m \rightarrow I_{w,r}}$

is the flow from the main image to the ultra-wide image. The intuition behind the consistency check is that, if an estimated flow is correct, then when a corresponding point is warped twice by the forward flow and backward flow, the output should be close to zero. To synthesize an AIF output, we not only need to find where occlusions occur but also to acquire clear results in those areas. We address them with an occlusion-aware synthesis network OAS-Net, as shown in Fig. 7.

We use an encoder-decoder structure with an efficient twist that replaces downsampling and upsampling operations with DWT and inverse DWT. DWT decomposes the feature map into high and low frequency contents, and IDWT reintegrates the frequency back into the feature map, which is similar to the DWT and IDWT in Fig. 6. Since we predict the fusion masks for AIF synthesis, the fusion mask is flat in most parts, resulting in less information in high frequency contents. Therefore the network can predict the fusion masks mainly in low frequency, then restore the masks to the original resolution with high frequency cues. In addition, the input resolution of a smartphone camera is high, we utilize wavelet transformation so that the mask can be predicted in lower resolution with less information loss than normal pooling layers.

As we predict the occluded areas, we simultaneously refine the corresponding blurred background with the help of homography-warped ultra-wide image $I_{w,r}$. We apply deformable convolution [53] to align the features of $I_{w,r}$ and I_m , then we reconstruct the refined image I_d based on the aligned features. Compared with previous works [5], [54], we predict deformable weights with the direct use of DWTs and inverse DWTs. We replace convolution with wavelet decomposition in deformable weights predictions because DWT is helpful in capturing global context and in preserving more information in downsampling than standard convolution or pooling. With the refined image I_d , the main image I_m , and the color-aligned ultra-wide $I_{w,c}$, we fuse them with their corresponding fusion masks M_d , M_m , and $M_{w,c}$ as

$$I_{AIF} = I_d \cdot M_d + I_m \cdot M_m + I_{w,c} \cdot M_{w,c}, \quad (2)$$

where I_{AIF} is the final AIF result. We apply a softmax function to limit the network training, so $M_{w,c} + M_d + M_m = 1$. M_d is supervised by M_c , because M_d mostly represents occlusions and low-confidence flow warping results usually occur in occluded regions.

D. Model Training

Loss Functions. For WDC-Net, we use the following loss:

$$\mathcal{L}_{WDC} = \mathcal{L}_1(I_{w,c} \cdot M_c, I_{gt} \cdot M_c) + \mathcal{L}_{ssim}(I_{w,c} \cdot M_c, I_{gt} \cdot M_c), \quad (3)$$

where L_1 is the ℓ_1 loss, and L_{ssim} is the structural similarity (SSIM) loss [55]. I_{gt} is the AIF ground truth image, and M_c is the confidence mask of optical flow computed from forward-backward consistency check. We use M_c to mitigate the influence of wrong warped regions.

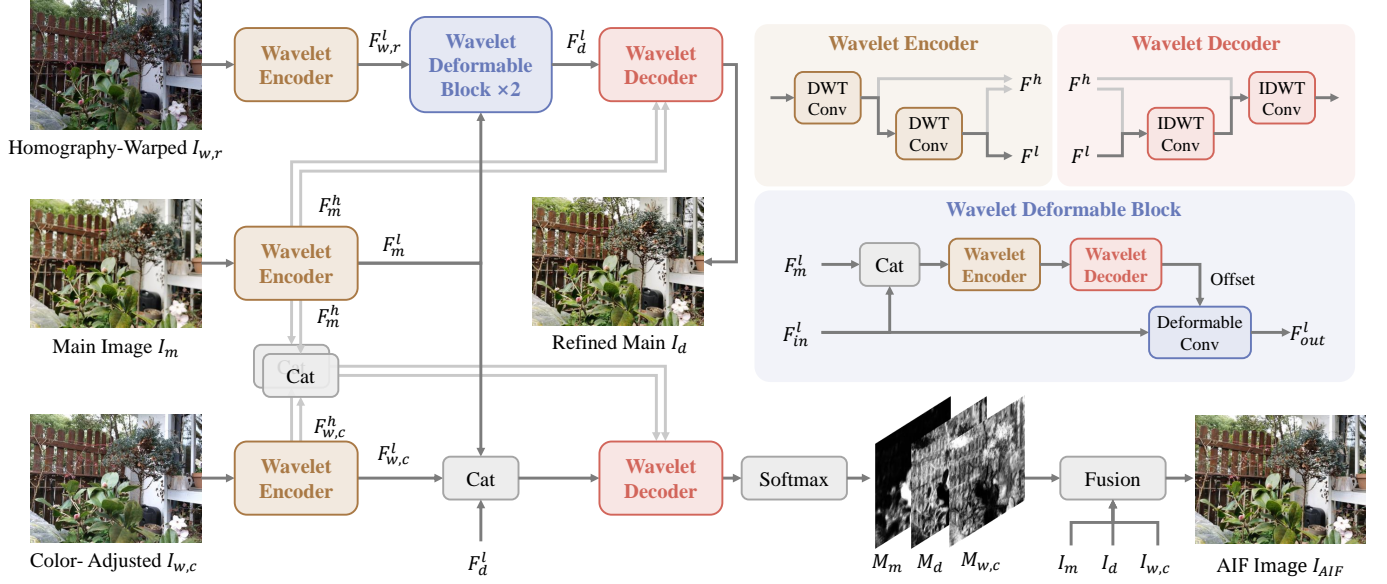


Figure 7. **Occlusion-aware synthesis network OAS-Net.** We modify the common encoder-decoder with wavelet transform, and we integrate the wavelet encoder and decoder into deformable convolution, where we predict the offset in the frequency domain. The three wavelet encoders of the three inputs ($I_{w,r}$, I_m , and $I_{w,c}$) share the same parameters. The two decoder branches output the fusion masks (M_m , M_d and $M_{w,c}$) for AIF synthesis and the refined main image I_d for occluded regions M_d .

For OAS-Net, we have the loss function

$$\begin{aligned} \mathcal{L}_{OAS} = & \mathcal{L}_1(I_{AIF}, I_{gt}) + \mathcal{L}_{sim}(I_{AIF}, I_{gt}) \\ & + \lambda \cdot \mathcal{L}_1(I_d, I_{gt}) + \lambda \cdot \mathcal{L}_{sim}(I_d, I_{gt}) \\ & + \delta \cdot \mathcal{L}_{bce}(M_m, M_{fuse}) \\ & + \delta \cdot \mathcal{L}_{bce}(M_d, (1 - M_{fuse}) \cdot M_c) \\ & + \delta \cdot \mathcal{L}_{bce}(M_w, (1 - M_{fuse}) \cdot (1 - M_c)), \end{aligned} \quad (4)$$

where \mathcal{L}_{bce} is the binary cross entropy loss. M_{fuse} is the fusion mask for generating AIF ground truth I_{gt} as

$$I_{gt} = M_{fuse} \cdot I_m^{fg} + (1 - M_{fuse}) \cdot I_m^{bg}, \quad (5)$$

where I_m^{fg} and I_m^{bg} are the main images focused on foreground and background, respectively. We describe the collections of I_m^{fg} , I_m^{bg} , M_{fuse} , and I_{gt} in Sec. IV. We use the binary cross entropy loss because we use a softmax function at the end of mask prediction, and the three mask ground truth images can all be defined as hard masks. $(1 - M_{fuse}) \cdot M_c$ is the ground truth of M_d because we want the refined image to fill in the low-confidence regions but leave out parts where the main image remains clear. $M_{fuse} \cdot (M_c - 1)$ is set as the ground truth of M_w because we want to ensure the three ground truth images add up to 1.

Implementation Details. We implement our model by PyTorch [56]. λ and δ are set to 0.5 and 0.1, respectively. We apply one DWT and one IDWT in WDC-Net. Two sets of DWT and IDWT are used in the encoder-decoder of OAS-Net, and three sets of DWT and IDWT are used in the dynamic offset estimator in the deformable convolution block. When training WDC-Net, we do not downsample the input beforehand. We downsample the image by a factor of 4 before training the OAS-Net. Both networks are trained for 40 epochs using the Adam optimizer [57]. WDC-Net uses a batch size of

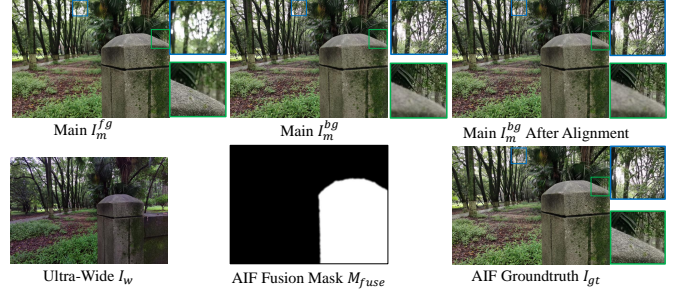


Figure 8. **An example of a set of training images.** The main images I_m^{fg} is focused on foreground, and I_m^{bg} is focused on background, as shown in the blue and green boxes. Since the focus breathing effect causes wider FOV in I_m^{bg} (the green box), we align I_m^{bg} to I_m^{fg} using flow warping. We also capture the ultra-wide image I_w , and the AIF ground truth image I_{gt} is generated from fusing the two main images.

8, and OAS-Net uses a batch size of 2. Our experiments are conducted on a single NVIDIA GeForce GTX 1080 Ti GPU, while some of the baseline experiments require an NVIDIA GeForce GTX 3090 GPU.

IV. DATASET COLLECTION

To train our AIF photo synthesis network, we collect the main/ultra-wide image pair for input, and the ground truth AIF image as supervision. We capture the smartphone AIF dataset from Huawei Mate 30 Pro, which uses an ultra-wide camera in its module. To obtain an image set, we first use the main camera to shoot two images; one focuses on the foreground, and one on the background. Then we switch to the ultra-wide camera to capture an image I_w following the default camera setting. We use a tripod to avoid camera shaking. We keep the exposure settings same for the two lenses, because we

intend to simulate the situation where the two cameras work simultaneously.

Now that we have three images per scene, we need to align the two main images due to focus breathing (a phenomenon that the FOV becomes narrow when focusing on a closer object). As shown in Fig. 8, the two main images I_m^{fg} and I_m^{bg} have view angle differences, where I_m^{bg} has a wider FOV. Therefore, we apply optical flow warping to align I_m^{bg} to I_m^{fg} . We take the two main camera images I_m^{fg} and the aligned I_m^{bg} as inputs and apply a multi-focus image fusion method [12] to synthesize the AIF ground truth I_{gt} for training. In addition, we acquire the fusion mask M_{fuse} for the two main images.

In all, we collect a dataset with 2686 indoor and outdoor scenes, with $2686 \times 2 = 5372$ image pairs. We split the dataset into 5000 training image sets and 372 evaluation image sets. Each set consists of I_m^{fg} , I_m^{bg} , I_w , M_{fuse} , and I_{gt} , as shown in Fig. 8. The ultra-wide image has a resolution of 3840×2592 , and other images have a resolution of 3648×2736 . The ultra-wide camera has a focal length of 18 mm and an aperture size of $f/1.8$, and the main camera has a focal length of 27 mm and an aperture size of $f/1.6$. To demonstrate the generalization of EasyAIF, we also capture 120 image pairs from iPhone13 to further solidify our claim.

V. RESULTS AND DISCUSSIONS

A. Baseline Approaches

We choose state-of-the-art baselines that are closely related to our new task, consisting of two types of approaches: 1) dual-pixel defocus deblurring: IFAN [9] and Restormer [11] and 2) dual-camera super-resolution: DCSR [7] and MASA-SR [6].

IFAN [9] is a single image defocus deblurring method which applies iterative adaptive convolutions. This model is trained on a dual-pixel dataset.

Restormer [11] addresses image restoration by using Transformer. The defocus deblurring model can be trained on a dual-pixel dataset.

DCSR [7] explores the dual camera super-resolution with aligned attention modules.

MASA-SR [6] uses matching acceleration and the spatial adaptation module to achieve reference-based super-resolution.

We use their pretrained models, and finetune the models on our dataset for a fair comparison.

B. Qualitative Results

We show the intermediate results of our approach in Figs. 9 to 11. In Fig. 9, we show the outputs of each components in EasyAIF, including the flow-warped ultra-wide image $I_{w,f}$, the color-adjusted ultra-wide image $I_{w,c}$, the refined main image I_d and their corresponding fusion masks. In Figs. 10 and 11, we show the high frequency and low frequency feature maps from wavelet transform in WDC-Net and OAS-Net.

We show qualitative comparisons with the baseline methods in Figs. 12 and 13. We demonstrate the results with different inputs: the main images focused on the foreground, and the main images focused on the background. One can observe: 1)

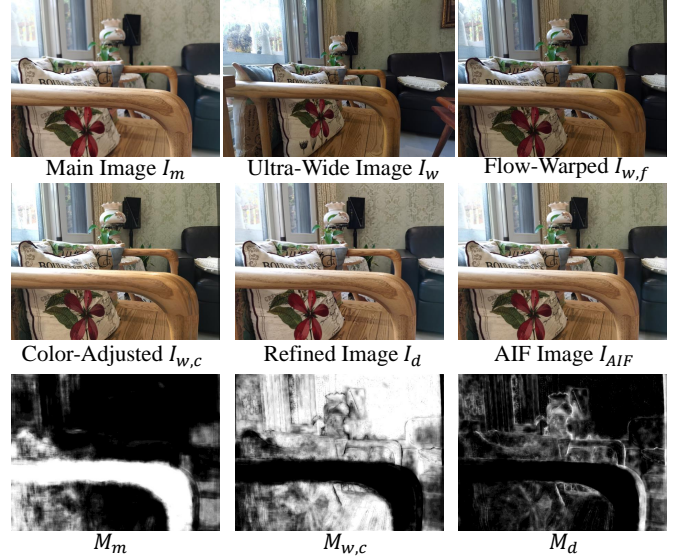


Figure 9. **Intermediate results of EasyAIF.** We demonstrate how the three components of EasyAIF work to synthesize I_{AIF} . The main image I_m is focused on the foreground regions. M_d includes the occlusions at the edges of foreground objects because the corresponding occluded regions of I_d are sharper than that of foreground-focused I_m .

DCSR [7] tends to oversmooth the blurred regions; 2) dual-pixel defocus deblurring methods IFAN [9] and Restormer [11] can effectively reduce the amount of blurring, but they fail to recover sharp details; 3) MASA-SR [6] performs well to restore some sharp details, but it still has many artifacts, such as the blue box in the first row; 4) Our method EasyAIF outperforms other baselines in generating AIF results.

C. Quantitative Results

As shown in Table I, we compare our method EasyAIF with the four baseline methods quantitatively on our synthesized smartphone AIF dataset. We split the evaluation dataset into two groups: the main images focused on foreground, and the main images focused on background. To measure the performance of different methods, we use LPIPS [58], PSNR, and SSIM as metrics. Results show that EasyAIF outperforms other methods numerically. Restormer [11] and DCSR [7] are the second best approaches. We demonstrate that EasyAIF is superior in terms of image focused on the foreground as well as on the background. Reference-based super-resolution methods MASA-SR [6] and DCSR perform well to super-resolve the already clear contents in the image, however, they fail to handle large defocus blur. Restormer and IFAN [9] are designed for defocus deblurring, but they do not fully utilize the sharp contents in the reference image I_w .

To prove the generalization across different devices, we capture 60 scenes on iPhone13 which also has an ultra-wide camera. The resolutions of the ultra-wide image and the main image are both 4032×3024 , which are different from those of Huawei Mate30 Pro. We evaluate our model on the iPhone dataset without finetuning. The results are shown in Table II. We show that EasyAIF still performs better than the SOTA SR/Deblur methods.

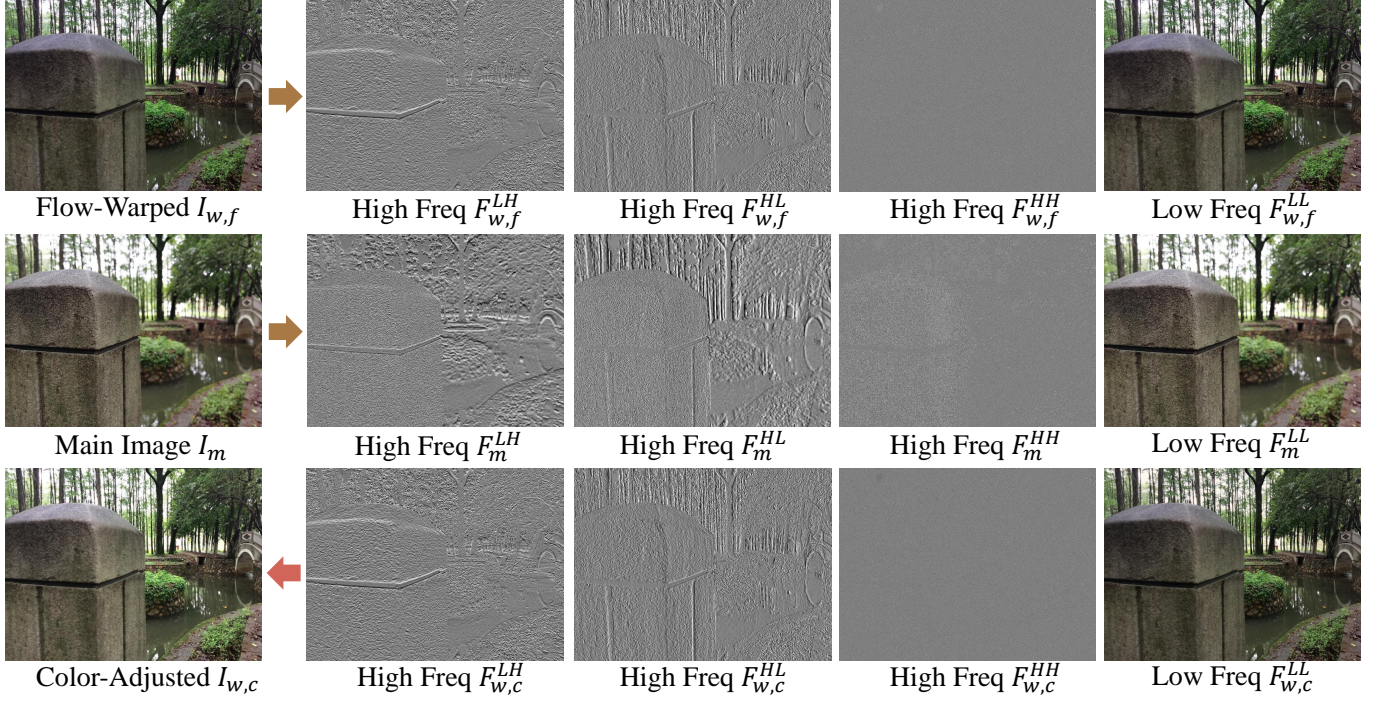


Figure 10. **Intermediate feature maps of WDC-Net.** The arrows indicate that $I_{w,f}$ and I_m are the inputs, and WDC-Net outputs $I_{w,c}$ from high frequency features and low frequency features. We can observe that the high frequency feature maps store information of the edges, and the low frequency feature maps relate more to the contents.

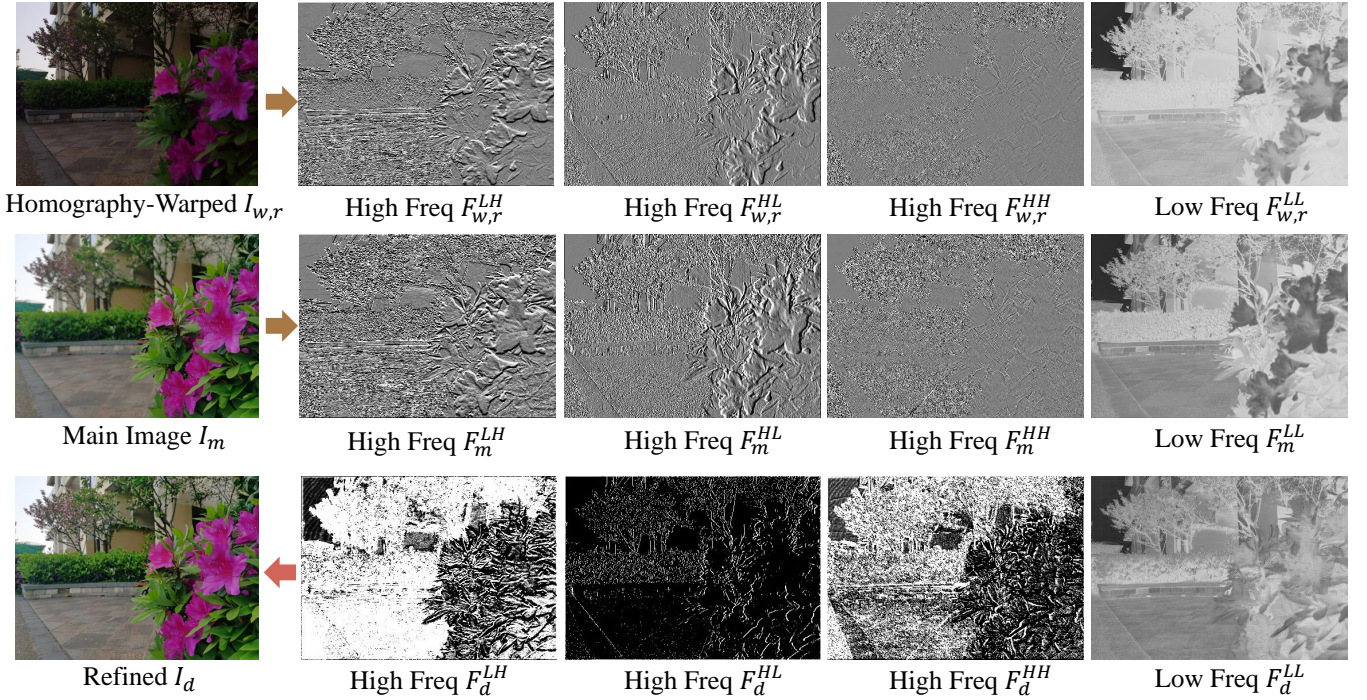


Figure 11. **Intermediate feature maps of the branch that produces the refined main image in OAS-Net.** The arrows indicate that $I_{w,r}$ and I_m are the inputs, and OAS-Net outputs I_d from high frequency features and low frequency features. We can observe that the high frequency feature maps store information of the edges, and the low frequency feature maps relate more to the contents.

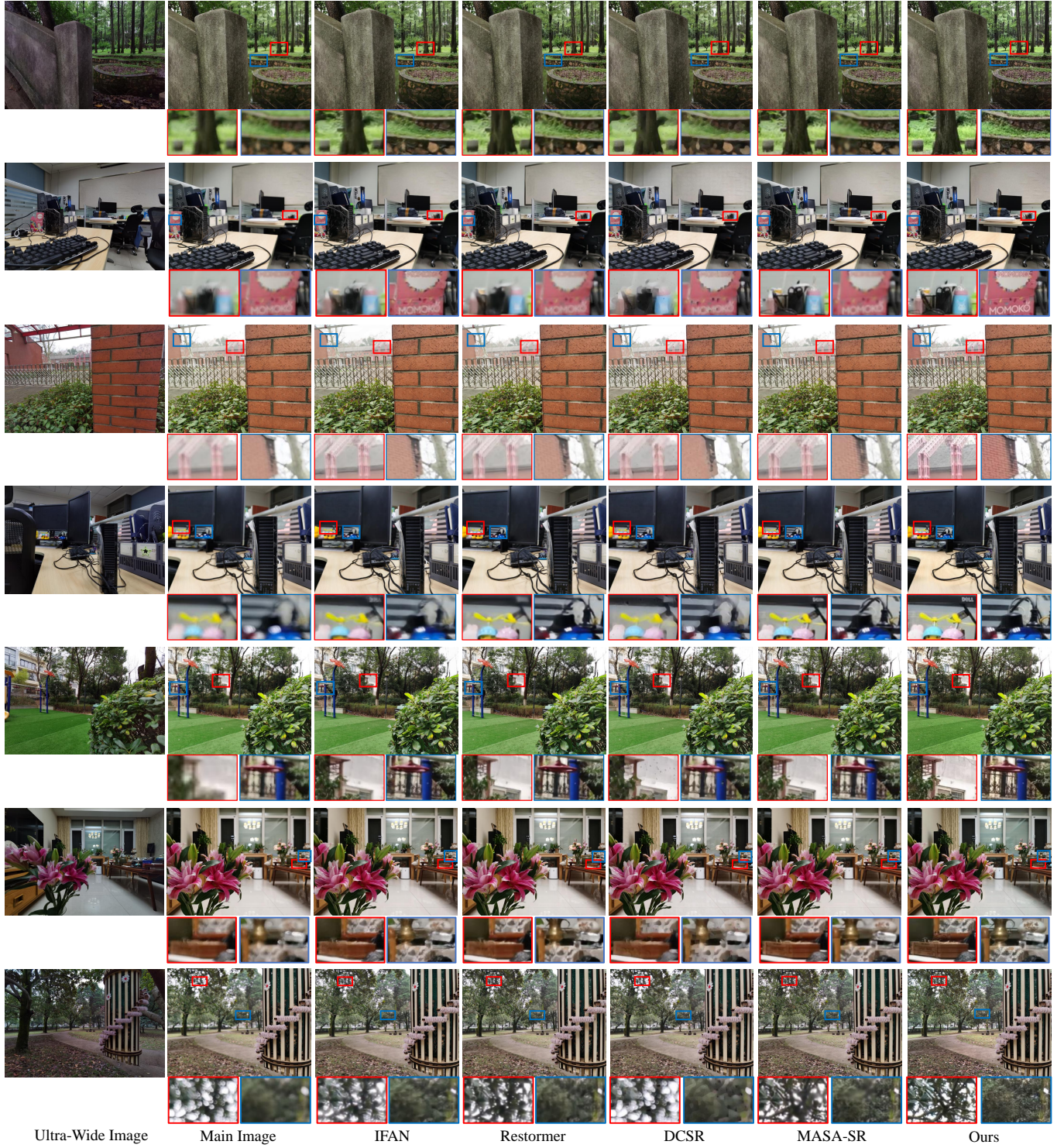


Figure 12. **Comparison with baselines on the smartphone AIF dataset.** The input pair is the main image focused on the foreground, and the ultra-wide image. Compared with the baselines, our proposed EasyAIF restores sharper details. Please zoom in to see the details.

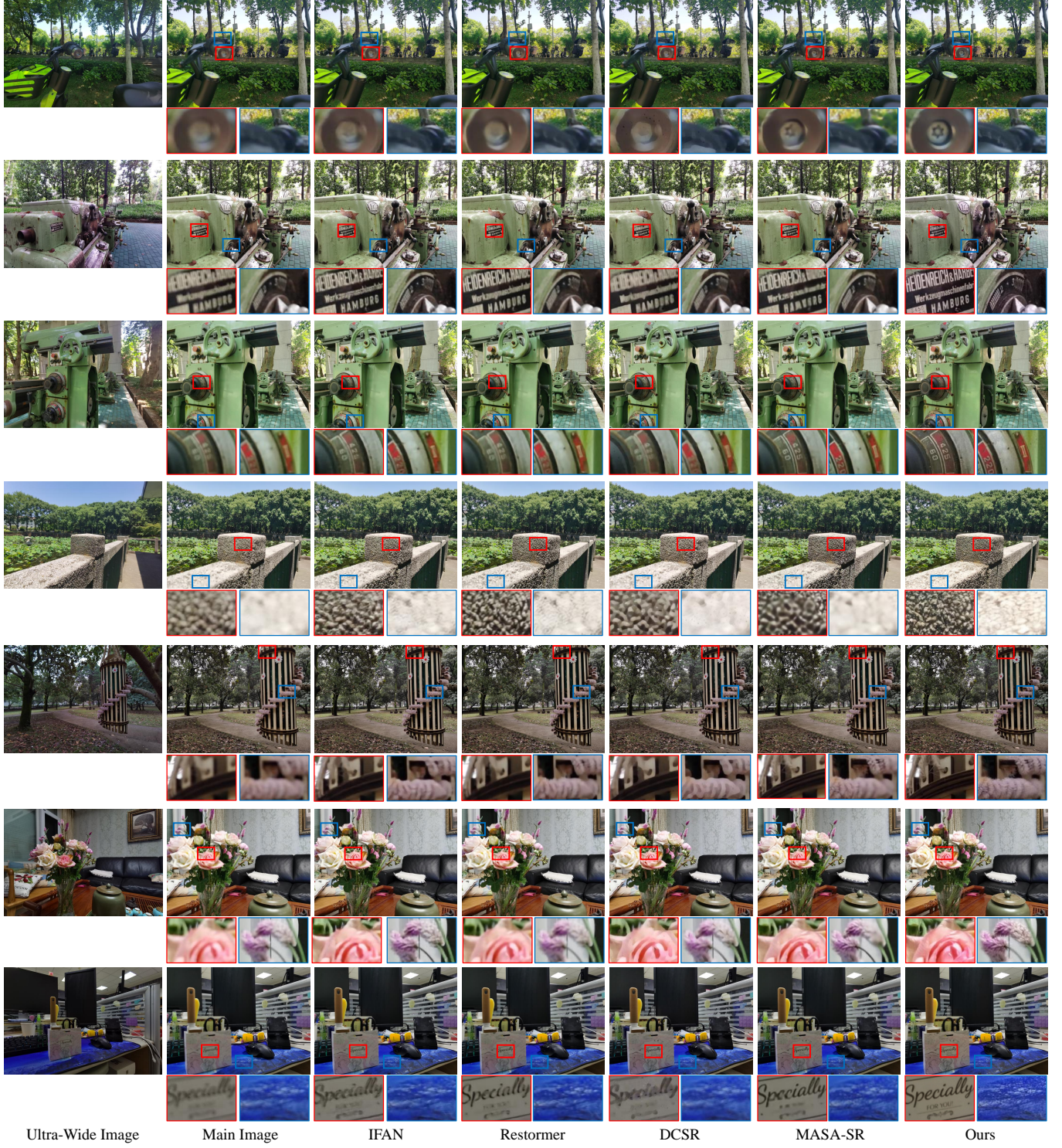


Figure 13. **Comparison with baselines on the smartphone AIF dataset.** The input pair is the main image focused on the background, and the ultra-wide image. Compared with the baselines, our proposed EasyAIF restores sharper details. Please zoom in to see the details.

TABLE I

QUANTITATIVE RESULTS ON OUR AIF DATASET. THE BEST PERFORMANCE IS IN **BOLDFACE**, AND THE SECOND BEST IS UNDERLINED.

Method	Focused on foreground			Focused on background			Total		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
IFAN [9]	25.09	0.742	0.402	28.38	0.835	0.281	26.73	0.789	0.341
Restormer [11]	<u>25.34</u>	<u>0.749</u>	0.380	28.74	0.840	0.270	<u>27.04</u>	0.795	0.325
DCSR [7]	24.11	0.722	<u>0.345</u>	<u>29.66</u>	<u>0.875</u>	<u>0.203</u>	26.88	<u>0.798</u>	0.274
MASA-SR [6]	25.24	0.717	0.422	25.72	0.722	0.419	25.48	0.719	0.421
EasyAIF (Ours)	25.75	0.828	0.185	30.58	0.921	0.130	28.16	0.874	0.158

TABLE II

QUANTITATIVE RESULTS ON IPHONE13 DATASET. THE BEST PERFORMANCE IS IN **BOLDFACE**, AND THE SECOND BEST IS UNDERLINED.

Method	Focused on foreground			Focused on background			Total		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
IFAN [9]	<u>23.48</u>	<u>0.638</u>	0.613	26.09	0.692	0.527	24.78	0.665	0.570
Restormer [11]	23.13	0.636	0.613	25.90	0.690	0.521	24.51	0.663	0.567
DCSR [7]	22.78	<u>0.638</u>	<u>0.479</u>	<u>26.80</u>	<u>0.728</u>	<u>0.398</u>	<u>24.79</u>	<u>0.683</u>	<u>0.438</u>
MASA-SR [6]	23.09	0.590	0.672	23.93	0.597	0.658	23.51	0.593	0.665
EasyAIF (Ours)	24.60	0.777	0.241	28.27	0.808	0.187	26.43	0.793	0.214

As shown in Table III, we also demonstrate the runtime of EasyAIF. In this table, “Others” means the remaining steps in image registration (descriptor extraction, correspondence matching, and consistency filtering), color alignment, and occlusion-aware synthesis. We evaluate the runtime on the image pair with the resolution of 4032x3024. It is worth mentioning that the SIFT keypoints extraction in image registration is time-consuming, and our total runtime is 1.59s on one NVIDIA 3090 GPU. The optimization of image registration is not our priority, so we simply use SIFT.

TABLE III

RUNTIME OF OUR METHOD. “OTHERS” MEANS THE REMAINING STEPS IN IMAGE REGISTRATION (DESCRIPTOR EXTRACTION, CORRESPONDENCE MATCHING, AND CONSISTENCY FILTERING), COLOR ALIGNMENT, AND OCCLUSION-AWARE SYNTHESIS.

	Keypoints Extraction	Others	Total
Time (s)	0.60	0.99	1.59

D. Ablation Study

We conduct ablation study on the framework components, including spatial alignment, color adjustment, and occlusion-aware synthesis. As shown in Table IV, our framework works best with all three modules. If we only use spatial alignment, the color discrepancy and the occlusion issue still exist. We mitigate these two problems by introducing WDC-Net and OAS-Net. To further prove our network design, we also conduct experiments on warping in spatial alignment, and ablation study on OAS-Net. As shown in Tab V, homography alignment and flow warping both improve our performance. Homography warping provides image-level alignment, and optical flow focuses on pixel-level warping. For occlusion-aware synthesis, as shown in Tab VI, we show that wavelet-based deformable refinement, as well as occlusion-aware image fusion, are both required to achieve better quantitative results. In addition, we demonstrate visual qualitative results on the ablation studies of OAS-Net and WDC-Net. In Fig. 14, we show that our WDC-Net is effective to align the color of flow-warped image $I_{w,f}$ to the main image I_m . For occlusion-aware

TABLE IV

ABLATION STUDY OF OUR AIF SYNTHESIS FRAMEWORK ON THE SMARTPHONE AIF DATASET. THE BEST PERFORMANCE IS IN **BOLDFACE**.

Framework Components			Total		
Spatial alignment	WDC-Net	OAS-Net	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
✓			16.06	0.675	0.342
✓	✓		20.65	0.732	0.304
✓		✓	21.65	0.739	0.174
✓	✓	✓	28.16	0.874	0.158

TABLE V

ABLATION STUDY ON WARPING OPERATIONS IN SPATIAL ALIGNMENT. THE BEST PERFORMANCE IS IN **BOLDFACE**.

Warping methods	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Homography	14.37	0.539	0.431
Flow	15.47	0.584	0.358
Homography+Flow	16.06	0.675	0.342

synthesis, as shown in Fig. 15, the color-adjusted $I_{w,c}$ suffers from occlusions, so we use OAS-Net to predict the occluded regions and refine the corresponding areas. Therefore, the final AIF result I_{AIF} performs better than I_m and $I_{w,c}$ on the occlusions.

TABLE VI

ABLATION STUDY ON OAS-NET. THE BEST PERFORMANCE IS IN **BOLDFACE**.

OAS structure	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Fusion	27.85	0.871	0.158
Refine	26.45	0.785	0.333
Refine+Fusion	28.16	0.874	0.158

E. User Study

To better evaluate the performance of EasyAIF, we conduct a user study on our smartphone AIF dataset. We collect 50 sets of images, and invite 20 people participating in this survey. We compare EasyAIF with each baseline separately and ask the participant to choose the more realistic result or choose none if it is hard to judge.

We build an online website for user study, the interface of the website is demonstrated in Fig. 16. “Input Image” is a defocused main camera image. “Focal Point” labels the target that is roughly refocused during the capturing. “Method 1” and “Method 2” display the results of two rendering methods, one of which is ours, and the other one is randomly selected from IFAN [9], Restormer [11], DCSR [7], and MASA-SR [6]. The positions of the two methods are also random. “Magnification Window” provides simultaneous local zoomed viewings for images in two rows. Users can utilize them to observe and compare the details of the two AIF images. When users are voting, if they cannot decide on the better result, each method gets 0.5 votes.

We show the comparison results in Table VII, where the number represents the preference of our approach over the

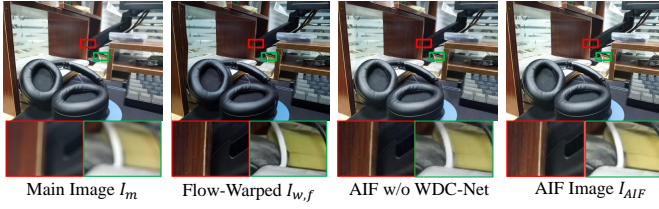


Figure 14. **Ablation study on our WDC-Net.** From the red and the green boxes, we can observe that before color alignment, flow warped image $I_{w,f}$ is darker than I_m , so the AIF result without WDC-Net also exhibits inconsistent color. On the other hand, I_{AIF} shows similar color and illumination to I_m .

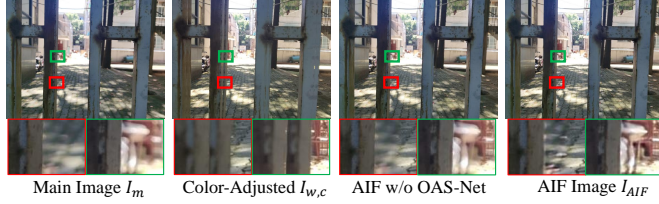


Figure 15. **Ablation study on our OAS-Net.** From the red and the green boxes, we show that the main image I_m and the color-adjusted $I_{w,c}$ produce blur or artifacts in occluded areas. I_{AIF} has sharper details than the AIF result without OAS-Net, and has fewer artifacts in occlusions than $I_{w,c}$.

other methods. One can observe that our approach is most favored.

TABLE VII

USER STUDY OF EASYAIF ON THE SMARTPHONE AIF DATASET. THE NUMBERS INDICATE THE PREFERENCE RATE OF OUR EASYAIF OVER OTHER APPROACHES.

Method	DCSR [7]	MASA-SR [6]	IFAN [9]	Restormer [11]
EasyAIF (Ours)	98.9%	87.1%	88.6%	85.8%

F. Comparisons with Multi-Focus Fusion Methods

Our method is different from multi-focus fusion methods [25], [26] from the inputs. Our method deals with two images from different cameras, which suffer from large misalignment. However, the input of the multi-focus fusion method is designed to be already aligned. Traditionally the focal stack images are shot by the same camera, so the captured images do not suffer from misalignment in space and color. As shown in Fig. 18, multi-focus fusion methods can



Figure 16. Interface of the user study website.



Figure 17. **Failure cases.** Our All-in-Focus (AIF) result may produce artifacts at edges due to large occlusions (row 1). The ultra-wide image may have defocused regions when the foreground object is too close to the camera, so the corresponding parts of the main image cannot be fixed (row 2).

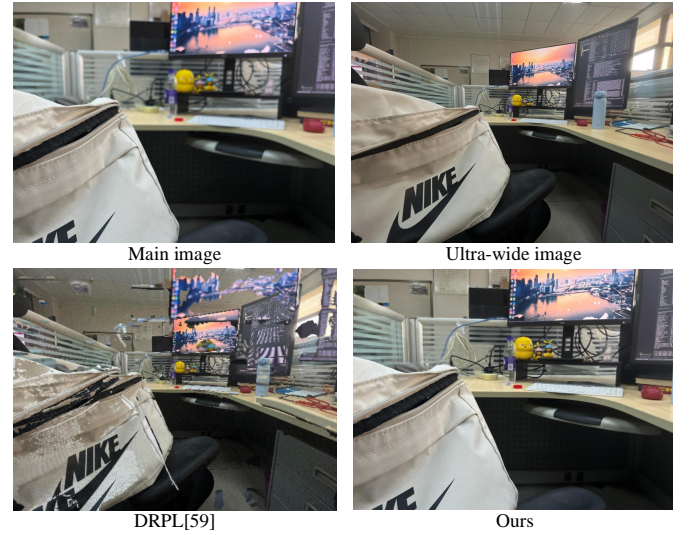


Figure 18. **Visualizations of the multi-focus fusion method and EasyAIF on iPhone13 dataset.** The multi-focus fusion method fails to fuse the two misaligned input images, while EasyAIF is a strong baseline to synthesize an AIF image from main/ultra-wide camera pair.

not deal with the misalignment between the main image and the ultra-wide image. In our smartphone AIF task, we use the main camera and the ultra-wide camera from a smartphone, so we need to take misalignment into considerations, which is challenging for our task. As shown in Table VIII, the quantitative result of the multi-focus fusion method is much inferior than the result of EasyAIF.

G. Failure Case Analyses

We show the failure cases in Fig. 17. The occlusion-aware synthesis network OAS-Net may produce artifacts if the occlusions are large. The fusion mask may include the wrongly warped occluded region of the ultra-wide image,

TABLE VIII

COMPARISONS OF THE MULTI-FOCUS FUSION METHOD AND EASYAIF ON IPHONE13 DATASET. THE BEST PERFORMANCE IS IN **BOLDFACE**.

Methods	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
DRPL [59]	11.28	0.386	0.933
EasyAIF	26.43	0.793	0.214

and the refined image sometimes cannot fix the blur from large occlusions. In addition, the ultra-wide image may fail to provide sharp details in extreme cases where the camera is too close to the object.

VI. CONCLUSION

We are the first to synthesize AIF photos from the main/ultra-wide camera pair. Compared with previous time-consuming methods, we have presented a point-and-shoot solution EasyAIF for AIF photography using a smartphone. We make use of the main/ultra-wide lens pair in modern smartphones to integrate both high-quality details from the main camera and sharp contents from the ultra-wide one. To align the two images, we use spatial warping for spatial alignment, and a wavelet dynamic network for color adjustment. To solve the occlusions brought by parallax, we propose an occlusion-aware synthesis network to refine the occluded regions and predict the fusion mask to generate AIF results. Results show that point-and-shoot AIF photo synthesis is viable from the main and ultra-wide camera pair. Although this framework works well in general, it still has some limitations, such as the inaccurate color transform due to spatial misalignment, and the boundary artifacts caused by imperfect refined results. We will address these issues in our future work.

REFERENCES

- [1] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen, "Interactive digital photomontage," in *ACM SIGGRAPH 2004 Papers*, 2004, pp. 294–302.
- [2] S. Pertuz, D. Puig, M. A. Garcia, and A. Fusiello, "Generation of all-in-focus images by noise-robust selective fusion of limited depth-of-field images," *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 1242–1251, 2012.
- [3] M. Lee and Y.-W. Tai, "Robust all-in-focus super-resolution for focal stack photography," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1887–1897, 2016.
- [4] H. Zheng, M. Ji, H. Wang, Y. Liu, and L. Fang, "Crossnet: An end-to-end reference-based super resolution network using cross-scale warping," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 88–104.
- [5] G. Shim, J. Park, and I. S. Kweon, "Robust reference-based super-resolution with similarity-aware deformable convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8425–8434.
- [6] L. Lu, W. Li, X. Tao, J. Lu, and J. Jia, "Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6368–6377.
- [7] T. Wang, J. Xie, W. Sun, Q. Yan, and Q. Chen, "Dual-camera super-resolution with aligned attention modules," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2001–2010.
- [8] A. Abuolaim and M. S. Brown, "Defocus deblurring using dual-pixel data," in *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2020, pp. 111–126.
- [9] J. Lee, H. Son, J. Rim, S. Cho, and S. Lee, "Iterative filter adaptive network for single image defocus deblurring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2034–2042.
- [10] H. Son, J. Lee, S. Cho, and S. Lee, "Single image defocus deblurring using kernel-sharing parallel atrous convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2642–2650.
- [11] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," *arXiv preprint arXiv:2111.09881*, 2021.
- [12] X. Qiu, M. Li, L. Zhang, and X. Yuan, "Guided filter-based multi-focus image fusion through focus region detection," *Signal Processing: Image Communication*, vol. 72, pp. 35–46, 2019.
- [13] Y. Chen, J. Guan, and W.-K. Cham, "Robust multi-focus image fusion using edge model and multi-matting," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1526–1541, 2017.
- [14] V. N. Gangapure, S. Banerjee, and A. S. Chowdhury, "Steerable local frequency based multispectral multifocus image fusion," *Information fusion*, vol. 23, pp. 99–115, 2015.
- [15] Q. Zhang, T. Shi, F. Wang, R. S. Blum, and J. Han, "Robust sparse representation based multi-focus image fusion with dictionary construction and local spatial consistency," *Pattern Recognition*, vol. 83, pp. 299–313, 2018.
- [16] K. Kodama and A. Kubota, "Efficient reconstruction of all-in-focus images through shifted pinholes from multi-focus images for dense light field synthesis and rendering," *IEEE Transactions on Image Processing*, vol. 22, no. 11, pp. 4407–4421, 2013.
- [17] S. Suwajanakorn, C. Hernandez, and S. M. Seitz, "Depth from focus with your mobile phone," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3497–3506.
- [18] J. Surh, H.-G. Jeon, Y. Park, S. Im, H. Ha, and I. So Kweon, "Noise robust depth from focus using a ring difference filter," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6328–6337.
- [19] P. Sakurikar and P. Narayanan, "Composite focus measure for high quality depth maps," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1614–1622.
- [20] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, pp. 191–207, 2017.
- [21] B. Xiao, B. Xu, X. Bi, and W. Li, "Global-feature encoding u-net (geu-net) for multi-focus image fusion," *IEEE Transactions on Image Processing*, vol. 30, pp. 163–175, 2020.
- [22] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502–518, 2020.
- [23] H. Zhang, Z. Le, Z. Shao, H. Xu, and J. Ma, "Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion," *Information Fusion*, vol. 66, pp. 40–53, 2021.
- [24] M. Maximov, K. Galim, and L. Leal-Taixé, "Focus on defocus: bridging the synthetic to real domain gap for depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1071–1080.
- [25] W. Zhao, D. Wang, and H. Lu, "Multi-focus image fusion with a natural enhancement via a joint multi-level deeply supervised convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1102–1115, 2018.
- [26] S. Liu, J. Chen, and S. Rahardja, "A new multi-focus image fusion algorithm and its efficient implementation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 5, pp. 1374–1384, 2019.
- [27] X. Luo, J. Peng, K. Xian, Z. Wu, and Z. Cao, "Bokeh rendering from defocus estimation," in *Computer Vision—ECCV 2020 Workshops*, 2020, pp. 245–261.
- [28] J. Peng, Z. Cao, X. Luo, H. Lu, K. Xian, and J. Zhang, "Bokehme: When neural rendering meets classical rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 16 283–16 292.
- [29] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," *ACM transactions on graphics (TOG)*, vol. 26, no. 3, pp. 70–es, 2007.
- [30] J. Park, Y.-W. Tai, D. Cho, and I. So Kweon, "A unified approach of multi-scale deep and hand-crafted features for defocus estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1736–1745.
- [31] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-laplacian priors," *Advances in neural information processing systems*, vol. 22, 2009.
- [32] J. Lee, S. Lee, S. Cho, and S. Lee, "Deep defocus map estimation using domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 222–12 230.
- [33] H. Kumar, S. Gupta, and K. Venkatesh, "Simultaneous estimation of defocus and motion blurs from single image using equivalent gaussian representation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3571–3583, 2019.

- [34] J. Shi, L. Xu, and J. Jia, "Just noticeable defocus blur detection and estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 657–665.
- [35] Y.-Q. Liu, X. Du, H.-L. Shen, and S.-J. Chen, "Estimating generalized gaussian blur kernels for out-of-focus image deblurring," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 829–843, 2020.
- [36] S. Xin, N. Wadhwa, T. Xue, J. T. Barron, P. P. Srinivasan, J. Chen, I. Gkioulekas, and R. Garg, "Defocus map estimation and deblurring from a single dual-pixel image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2228–2238.
- [37] A. Abuolaim, M. Afifi, and M. S. Brown, "Improving single-image defocus deblurring: How dual-pixel images help through multi-task learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1231–1239.
- [38] A. Abuolaim, M. Delbracio, D. Kelly, M. S. Brown, and P. Milanfar, "Learning to reduce defocus blur by realistically modeling dual-pixel data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2289–2298.
- [39] L. Pan, S. Chowdhury, R. Hartley, M. Liu, H. Zhang, and H. Li, "Dual pixel exploration: Simultaneous depth estimation and image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4340–4349.
- [40] L. Ruan, B. Chen, J. Li, and M.-L. Lam, "Aifnet: All-in-focus image restoration network using a light field-based dataset," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 675–688, 2021.
- [41] Y. Zhou, C. Barnes, E. Shechtman, and S. Amirghodsi, "Transfill: Reference-guided image inpainting by merging multiple color and spatial transformations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2266–2276.
- [42] C. Luo, Y. Li, K. Lin, G. Chen, S.-J. Lee, J. Choi, Y. F. Yoo, and M. O. Polley, "Wavelet synthesis net for disparity estimation to synthesize dslr calibre bokeh effect on smartphones," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2407–2415.
- [43] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157.
- [44] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, "Working hard to know your neighbor's margins: Local descriptor learning loss," *Advances in neural information processing systems*, vol. 30, 2017.
- [45] C. Zhao, Z. Cao, C. Li, X. Li, and J. Yang, "Nm-net: Mining reliable neighbors for robust feature correspondences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 215–224.
- [46] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [47] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2020, pp. 402–419.
- [48] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–12, 2017.
- [49] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.
- [50] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 030–11 039.
- [51] J. Liang, H. Zeng, and L. Zhang, "High-resolution photorealistic image translation in real-time: A laplacian pyramid translation network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9392–9400.
- [52] Q. Chen and V. Koltun, "Full flow: Optical flow estimation by global optimization over regular grids," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4706–4714.
- [53] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 764–773.
- [54] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "Tdan: Temporally-deformable alignment network for video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3360–3369.
- [55] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Computational Imaging*, vol. 3, no. 1, pp. 47–57, 2017. [Online]. Available: <https://doi.org/10.1109/TCI.2016.2644865>
- [56] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Advances in Neural Information Processing Systems Workshops (NIPSW)*, 2017.
- [57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations (ICLR)*, 2014.
- [58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
- [59] J. Li, X. Guo, G. Lu, B. Zhang, Y. Xu, F. Wu, and D. Zhang, "Drpl: Deep regression pair learning for multi-focus image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4816–4831, 2020.



Xianrui Luo received the B.S. degree from Huazhong University of science and Technology, Wuhan, China, in 2020. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China.

His research interests include image restoration and computational photography.



Juewen Peng received the B.S. degree from Huazhong University of science and Technology, Wuhan, China, in 2020. He is currently pursuing the M.S. degree in the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China.

His research interests include low-level vision, computational photography, bokeh rendering, deblurring, and image generation.



Hao Lu received the Ph.D. degree from Huazhong University of Science and Technology, Wuhan, China, in 2018.

He was a Postdoctoral Fellow with the School of Computer Science, The University of Adelaide, Australia. He is currently an Associate Professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China. His research focuses on dense prediction problems in computer vision.



Weiyue Zhao received the B.S. degree from Huazhong University of science and Technology, Wuhan, China, in 2020. He is currently pursuing the M.S. degree with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China.

His research interests include computer vision and machine learning, with particular emphasis on image registration, multi-view stereo and various computer vision applications in video.



Ke Xian is a research fellow at S-Lab, Nanyang Technological University (NTU), Singapore. He got his Ph.D. degree at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology (HUST), China.

His research interests primarily centers on algorithms issues in computer vision and deep learning, including depth estimation from single images, semantic image segmentation and 2D-to-3D conversion.



Zhiguo Cao (Member, IEEE) received the B.S. and M.S. degrees in communication and information system from the University of Electronic Science and Technology of China and the Ph.D. degree in pattern recognition and intelligent system from Huazhong University of Science and Technology.

He is currently a Professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. His research interests spread across image understanding and analysis, depth information extraction, 3d video processing, motion detection, and human action analysis. He has published dozens of papers at international journals and prominent conferences, which have been applied to automatic observation system for crop growth in agricultural, for weather phenomenon in meteorology and for object recognition in video surveillance system based on computer vision.