DAVANet: Stereo Deblurring with View Aggregation

Shangchen Zhou[1] Jiawei Zhang[1] Wangmeng Zuo[2*] Haozhe Xie[2] Jinshan Pan[3] Jimmy Ren[1]
[1]SenseTime Research [2]Harbin Institute of Technology, Harbin, China
[3]Nanjing University of Science and Technology, Nanjing, China
https://shangchenzhou.com/projects/davanet

Abstract

*Nowadays stereo cameras are more commonly adopted in emerging devices such as dual-lens smartphones and unmanned aerial vehicles. However, they also suffer from blurry images in dynamic scenes which leads to visual discomfort and hampers further image processing. Previous works have succeeded in monocular deblurring, yet there are few studies on deblurring for stereoscopic images. By exploiting the two-view nature of stereo images, we propose a novel stereo image deblurring network with **D**epth **A**wareness and **V**iew **A**ggregation, named **DAVANet**. In our proposed network, 3D scene cues from the depth and varying information from two views are incorporated, which help to remove complex spatially-varying blur in dynamic scenes. Specifically, with our proposed fusion network, we integrate the bidirectional disparities estimation and deblurring into a unified framework. Moreover, we present a large-scale multi-scene dataset for stereo deblurring, containing 20,637 blurry-sharp stereo image pairs from 135 diverse sequences and their corresponding bidirectional disparities. The experimental results on our dataset demonstrate that DAVANet outperforms state-of-the-art methods in terms of accuracy, speed, and model size.*

1. Introduction

With the wide use of dual-lens smartphones, unmanned aerial vehicles and autonomous robots, stereoscopic vision has attracted increasing attention from researchers. Relevant studies not only covers traditional stereo tasks, such as stereo matching [42, 3, 29] and scene flow estimation [22, 23, 11], but also some novel tasks for improving visual effects of stereoscopic 3D contents, for example, stereo super-resolution [12], stereo video retargeting [18] and stereo neural style transfer [4, 7]. However, stereo image deblurring has rarely been discussed. In fact, the images captured by handheld or on-board stereo cameras often contain blur due to camera shake and object motion. The blurry stereo images would cause visual discomfort to viewers and make it difficult for further image processing.

*Corresponding author

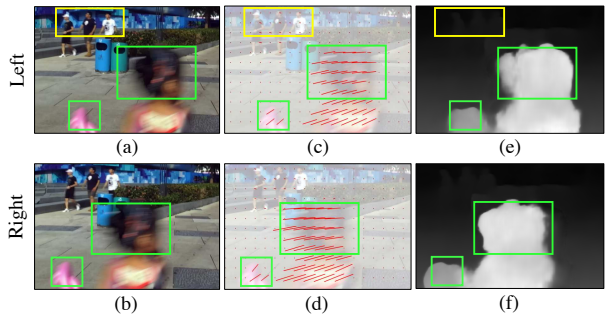


Figure 1: Depth-varying and view-varying blur. (a, b) are the stereo blurry images, (c, d) are the motion trajectories in terms of optical flow which models the blur kernels and (e, f) are the estimated disparities. The objects with different depths have different blurs which can be seen between green and yellow boxes. In addition, the green boxes show that the blurs are different between two views. The proposed *DAVANet* makes use of the above properties for deblurring.

Dynamic scene deblurring from a single blurry image is a highly ill-posed task. Due to depth variation and object/camera motion in dynamic scenes, it is difficult to estimate spatially variant blur with the limited information from single observation. Although the existing CNN based methods [38, 44, 16, 2, 24, 35] have achieved encouraging results in monocular deblurring, they still fail when handling complicated non-uniform blur. To the best of our knowledge, there are few traditional methods [40, 33, 28] proposed to exploit stereo information for deblurring, where a coarse depth or piecewise rigid 3D scene flow is utilized to estimate blur kernels in a hierarchical or iterative framework. However, they are time-consuming due to the complex optimization process.

With a stereopsis configuration, our motivation is based on two observations: (i) Depth information can provide helpful prior information for estimating spatially-varying blur kernels. The near points are more blurry than the distant ones in a static scene which can be seen between the green and yellow boxes in Figure 1. Compared monocular-based algorithms, the proposed stereo-based method can

obtain more accurate depth information by the disparity estimation. (ii) The varying information in corresponding pixels cross two stereo views can help blur removal. In Section 3.1, we demonstrate that the corresponding pixels in two different views have different blurs due to the motion perpendicular towards the camera and rotation, which is shown as the green boxes in Figure 1. The network can benefit from aggregated information, where the sharper pixel can be transferred and selected by using an adaptive fusion scheme. Two views can also share varying information, e.g., non-occlusion areas, caused by different viewpoints.

Inspired by these two insights, we propose a novel depth-aware and view-aggregated stereo deblurring network, named *DAVANet*. It consists of *DeblurNet* and *DispBiNet*, for image deblurring and bidirectional disparities estimation respectively. The *DeblurNet* and the *DispBiNet* are integrated at feature domain by the proposed fusion network, named *FusionNet*. Specifically, the *DispBiNet* provides depth-integrated features and bidirectional disparities for the *FusionNet*. The *FusionNet* fully exploits these inputs and enriches the *DeblurNet* features with embedding depth and the other view information. With the perception of 3D scene information from stereo images, the proposed method is effective for dynamic scene deblurring. Finally, to obtain richer contextual information, a context module is designed to incorporate the multi-scale contextual information by applying several parallel atrous convolutions with different dilation rates.

Currently, there is no particular dataset for stereo deblurring. As a result, we propose a large-scale multi-scene stereo blurry image dataset. It consists of 20,637 blurry-sharp stereo image pairs from 135 different sequences (98 for training and 37 for testing) and corresponding bidirectional disparities obtained from the ZED stereo camera [1]. We adopt the blur generation method used in [20, 24, 35], that is, approximating a longer exposure by accumulating the frames in an image sequence. We first interpolate frame of captured videos to a very high frame rate (480 fps) using frame interpolation method proposed in [25] and then average the sharp sequence to create a blurry image.

The main contributions are summarized as follows:

- We propose a unified network for stereo deblurring. The *DispBiNet* predicts the bidirectional disparities for depth awareness as well as view information aggregation in the *FusionNet*, which helps the *DeblurNet* to remove dynamic scene blur from stereo images.
- We present a first large-scale multi-scene dataset for stereo deblurring, which consists of 20,637 stereo images from 135 diverse scenes. It is currently the largest dataset for deblurring.
- We both quantitatively and qualitatively evaluate our method on our dataset and show that it performs favorably against state-of-the-art algorithms in terms of accuracy, speed as well as model size.

2. Related Work

Our work is a new attempt for solving stereo image deblurring by integrating blur removal and disparity estimation into a unified network. The following is a review of relevant works on monocular single-image deblurring, monocular multi-image deblurring, as well as stereo image deblurring respectively.

Single-image Deblurring. Many methods have been proposed for single-image deblurring. Some natural image priors are designed to help blur removal, such as L_0-regularized prior [41], dark channel prior [27], and discriminative prior [19]. However, it is difficult for these methods to model spatially variant blur in dynamic scenes. To model the non-uniform blur, some depth-based methods [17, 31, 9, 30] that utilize the predicted depth map to estimate different blur kernels. When the blur kernels are not be accurately estimated, they tend to generate visual artifacts in restored images. Moreover, they are computationally inefficient due to the complex optimization process.

Recent years have witnessed significant advances in single image deblurring by CNN-based models. Several methods [36, 6] use CNNs to estimate the non-uniform blur kernels. A conventional non-blind deblurring algorithm [45] is used removing blur, which is time-consuming. More recently, many end-to-end CNN models for image deblurring have also been proposed [24, 26, 43, 38, 44, 16]. To obtain a large receptive field in the network for blur removal, [38] and [38] develop a very deep multi-scale networks in coarse-to-fine manner. Different from [24], Tao *et al.* [38] share the weights of the network at three different spatial scales and use the LSTM to propagate information across scales. To handle spatially variant blur in dynamic scenes, Zhang *et al.* [44] adopt a VGG network to estimate the pixel-wise weights of the spatially variant RNNs [21] for blur removal in feature space. Noroozi *et al.* [26] build skip connections between the input and output, which reduces the difficulty of restoration and ensures color consistency. In addition, the adversarial loss is used in [24, 16] to restore more texture details.

Multi-image Deblurring. Recently, several CNN-based methods [35, 10, 14, 2] have been proposed for monocular multi-image (video/burst) deblurring. [35] and [14] align the nearby frames with the reference frame to restore the sharp images, which can obtain more rich information cross different images. Kim *et al.* [10] propose a frame recurrent network to aggregate multi-frame features for video deblurring. By repeatedly exchanging the features across the burst images, Aittala *et al.* [2] propose an end-to-end burst deblurring network in an order-independent manner. Based on the observations that the different images from video or burst are blurred differently, these multi-image fusion methods usually lead to good performance.

Stereo Deblurring. So far, there are few traditional methods [40, 33, 28] that leverage the scene information (i.e.,

disparity and flow) from stereo images for deblurring. Xu and Jia [40] partition the image into regions according to disparity (depth) estimated from stereo blurry images and estimate their blur kernels hierarchically. The methods [33, 28] propose a stereo video deblurring framework, where 3D scene flow estimation and blur removal are conducted jointly so that they can enhance each other with an iterative manner.

3. Proposed Method

3.1. Motivation

The motivation that utilizing stereo camera for dynamic scene deblurring is inspired by two observations, which is exemplified in Figure 1. First, we find that nearby object points are more blurry than distant ones and stereo cameras can provide depth information (disparity). Second, the two views of the stereo camera may produce different sizes of the blur to the same object because of relative motion along the depth direction and camera rotation. The sharper view can help the other view to restore better by sharing its information. In this section, we analyze the above observations in details with the assumption that the stereo camera has already been rectified.

Depth-Varying Blur. In [40], Xu and Jia have analyzed the relationship between blur size and depth. In Figure 2(a), we simply restate it by only considering the relative translation parallel to the image plane I. According to the similar triangles theorem:

$$\Delta X / \Delta P = f / z, \tag{1}$$

in which ΔX, ΔP, f and z denote the size of blur, the motion of object point, focal length, and depth of object point, respectively. Eq. 1 shows that blur size ΔX is inversely proportional to depth z if motion ΔP is fixed, which means that the closer object will generate the larger blur.

View-Varying Blur. For the stereo setups, the relative movements between the object point P and two lens of stereo camera are different because the point P is captured from different viewpoints. These differences make the object exhibit different blurs under the two views. Here, we consider two scenarios: relative translation along depth direction and rotation. For translation, we assume the object point P moves from P_t to P_{t+1} along the depth direction in Figure 2(b). According to the similar triangles theorem:

$$\Delta X_L / \Delta X_R = \overline{P_t M} / \overline{P_t N} = h / (h + b), \tag{2}$$

where b is the baseline of the stereo camera and h is the distance between left camera C_L and line $\overline{P_t P_{t+1}}$. It demonstrates that the blur sizes for two views of a stereo camera are different due to relative translation in depth direction.

As to relative rotation in Figure 2(c), the velocities of two lens v_{C_L}, v_{C_R} of the stereo camera are proportional to

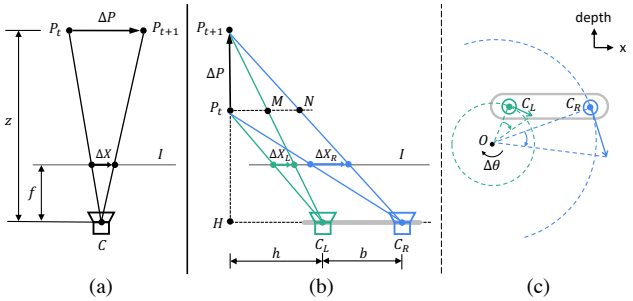


Figure 2: (a) is the depth-varying blur due to relative translation parallel to the image plane. (b) and (c) are the view-varying blur due to relative translation along depth direction and rotation. Note that all complex motion can be divided into above three relative sub-motion patterns.

the corresponding radiuses of the rotation $\overline{C_L O}$, $\overline{C_R O}$:

$$v_{C_L} / v_{C_R} = \overline{C_L O} / \overline{C_R O}. \tag{3}$$

In addition, the directions of the velocities are different due to relative rotation. As a result, both the size and direction of the blur vary between two views. The proposed network can utilize the information from the clearer view to help restore a better image for the more blurry one.

3.2. Network Architecture

The overall pipeline of the proposed *DAVANet* is illustrated in Figure 3. It consists of three sub-networks: *DeblurNet* for single-image deblurring, *DispBiNet* for bidirectional disparities estimation and *FusionNet* for fusing depth and two-view informations in an adaptive selection manner. Note that we adopt small convolution filters (3×3) to construct these three sub-networks and find that using the large filters does not significantly improve the performance.

DeblurNet. The U-Net based structure of *DeblurNet* is shown in Figure 4(a). We use the basic residual block as the building block, which has been proved effectiveness in deblurring [24, 38]. The encoder outputs features with $\frac{1}{4} \times \frac{1}{4}$ of the input size. Afterward, the following decoder reconstructs the sharp image with full resolution via two upsampled residual blocks. The skip-connections between corresponding feature maps are used between encoder and decoder. In addition, we also adopt a residual connection between the input and output. which makes it easy for the network to estimate the residual between blurry-sharp image pair and maintains color consistency.

To enlarge the receptive field and obtain the multi-scale information, the scale-recurrent scheme is popularly adopted in [24, 38]. Despite their performance improvement, they greatly increase the complexity of time and space. To solve this, we employ the two atrous residual blocks and a *Context Module* between encoder and decoder to obtain richer features. The *Context module* will be de-

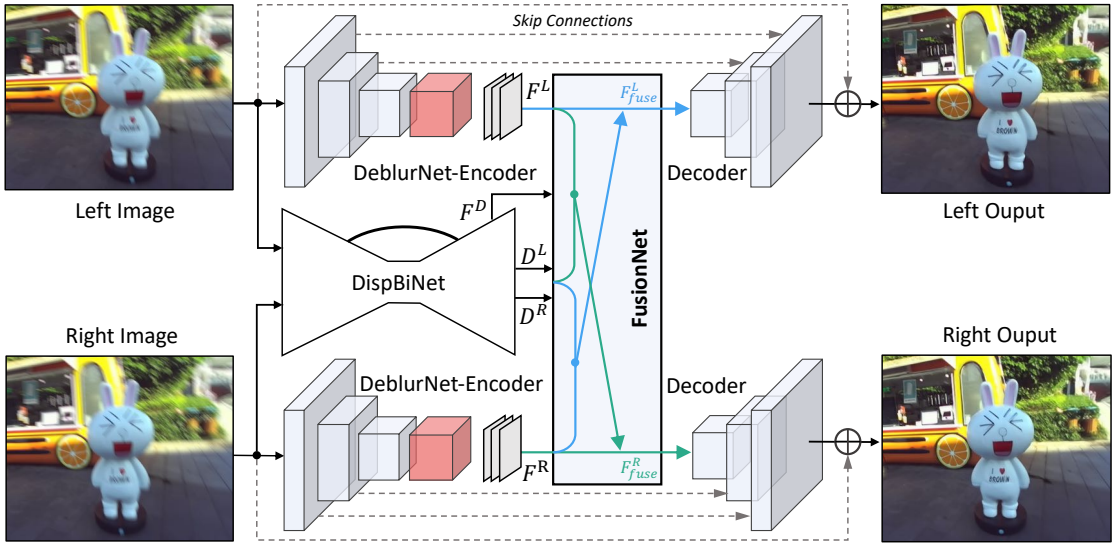


Figure 3: The overall structure of stereo deblurring network *DAVANet*, where the depth and the two-view information from the *DispBiNet* and the *DeblurNet* are integrated in *FusionNet*. Note that the *DeblurNet* shares weights for two views.

scribed in later a section. It should be noted that the *DeblurNet* uses shared weights for both views.

DispBiNet. Inspired by DispNet [22] structure, we propose a small *DispBiNet* as shown in Figure 4(b). Different from DispNet, the proposed *DispBiNet* can predict bidirectional disparities in one forward process. The bidirectional prediction has been proved better than unidirectional prediction in scene flow estimation [11]. The output is the full resolution with three times downsample and upsample in this network. In addtion, the residual block, atrous residual block, and context module are also used in *DispBiNet*.

Context Module. To embed the multi-scale features, we propose the *Context Module* (a slightly modified version of ASPP [5]) for *DeblurNet* and *DispBiNet*, which contains parallel dilated convolutions with different dilated rates, as show in Figure 4. The four dilated rates are set to: $1, 2, 3, 4$. *Context Module* fuses richer hierarchical context information that benefit both blur removal and disparity estimation.

Fusion Network. To exploit depth and two-view information for deblurring, we introduce the fusion network *FusionNet* to enrich the features with the disparities and the two views. For simplicity, we take left image as reference in this sections. As shown in Figure 5, *FusionNet* takes the original stereo images I^L, I^R, the estimated disparity of left view D^L, features F^D of the second last layer of *DispBiNet* and features F^L, F^R from *DeblurNet* encoder as input in order to generate the fused features F^L_{fuse}.

For two-view aggregation, the estimated left-view disparity D^L is used to warp right-view features F^R of *DeblurNet* to the left view, denoted as $W^L(F^R)$. Instead of directly concatenating $W^L(F^R)$ and F^L, the sub-network *GateNet* is employed to generate a soft gate map G^L ranging from 0 to 1. The gate map can be utilized to fuse features F^L and $W^L(F^R)$ in an adaptive selection scheme, that is, it selects helpful features and rejects incorrect ones from the other view. For example, at occlusion or false disparity regions, the values in the gate map tend to be 0, which suggest that only the features of reference view F^L should be adopted. The *GateNet* consists of five convolutional layers as shown in Figure 5. Its input is absolute difference of input left image I^L and the warped right image $W^L(I^R)$, namely $\left|I^L - W^L(I^R)\right|$, and the output is a single channel gate map. All feature channels share the same gate map to generate the aggregated features:

$$F^L_{views} = F^L \odot (1 - G^L) + W^L(F^R) \odot G^L, \quad (4)$$

where $\odot$ denotes element-wise multiplication.

For depth awareness, a sub-network *DepthAwareNet* containing three convolutional layers is employed, and note that this sub-network is not shared by both views. Given the disparity D^L and the second last layer features F^D of *DispBiNet*, *DepthAwareNet-left* produces the depth-involved features F^L_{depth}. In fact, *DepthAwareNet* learns the depth-aware prior implicitly, which helps for dynamic scene blur removal.

Finally, we concatenate the original left-view features F^L, view-aggregated features F^L_{views}, and depth-aware features F^L_{depth} to generate the fused left-view features F^L_{fuse}. And then, we feed the F^L_{fuse} to the decoder of *DeblurNet*. Note that the fusion processings of two views are the same.

3.3. Losses

Deblurring Losses. For Deblurring, we consider two loss functions to measure the difference between the restored image $\hat{I}$ and sharp image I for both two views L, R. The

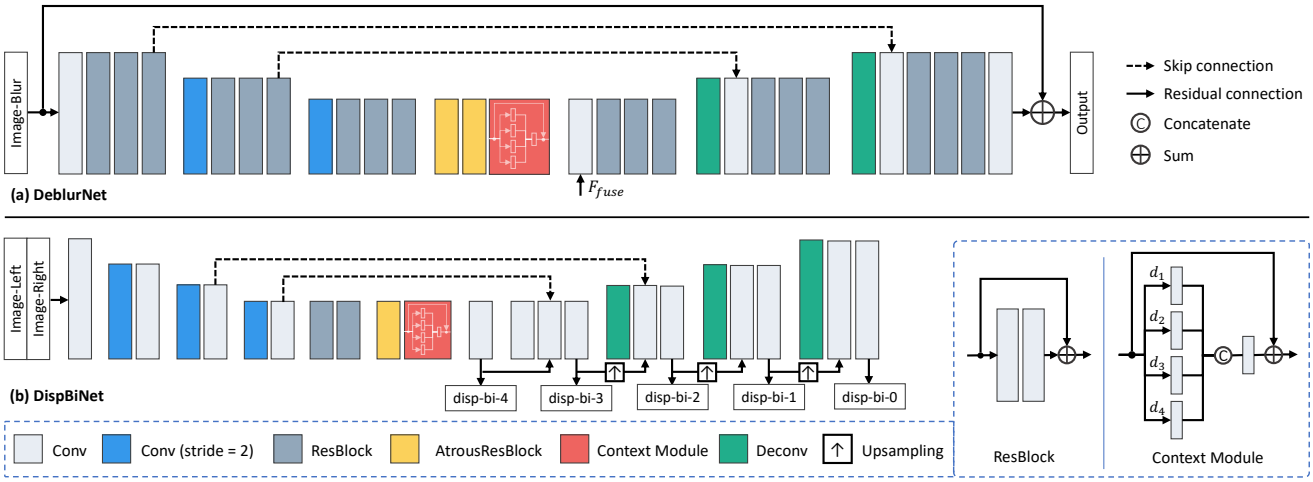


Figure 4: The detail structures of *DeblurNet* and *DispBiNet*. To get richer multi-scale features, the *Context Module* is adopted in both *DeblurNet* and *DispBiNet*, which contains parallel dilated convolutional layers with different dilation rates.

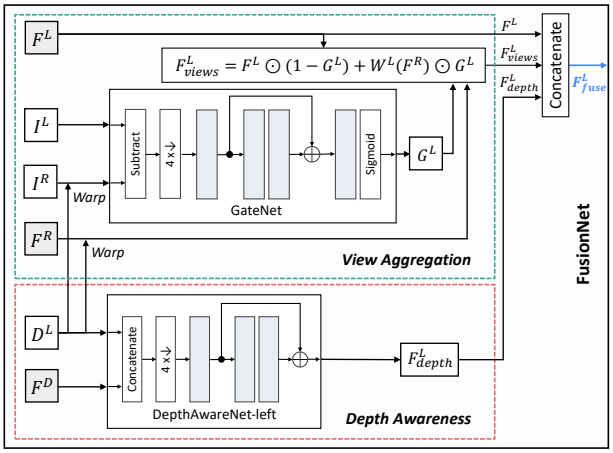


Figure 5: Fusion network. The *FusionNet* consists two components: depth awareness and view aggregation, which generate the depth-view fused feature for the decoder of *DeblurNet*. For simplicity, we only show the forward pass for the left image.

first loss is MSE loss:

$$\mathcal{L}_{mse} = \frac{1}{2CHW} \sum_{k \in \{L,R\}} ||\hat{I}^k - I^k||^2, \tag{5}$$

where C, H, W are dimensions of image. The other loss function is perceptual loss proposed in [13], which is defined as the l_2-norm between the VGG-19 [34] features of restored image $\hat{I}$ and sharp image I:

$$\mathcal{L}_{perceptual} = \frac{1}{2\mathcal{C}_j \mathcal{H}_j \mathcal{W}_j} \sum_{k \in \{L,R\}} ||\Phi_j(\hat{I}^k) - \Phi_j(I^k)||^2, \tag{6}$$

where $\mathcal{C}_j, \mathcal{H}_j, \mathcal{W}_j$ are dimensions of the features, and $\Phi_j(\cdot)$ denotes the features from the j-th convolution layer within the pretrained VGG-19 network. In our work we use the features from conv3-3 layer (j=15). The overall loss function for deblurring is:

$$\mathcal{L}_{deblur} = \sum_{k \in \{L,R\}} w_1 \mathcal{L}^k_{mse} + w_2 \mathcal{L}^k_{perceptual}, \tag{7}$$

where the weights w_1, w_2 of two losses are set to $1, 0.01$ in our experiments, respectively.

Disparity Estimation Loss. For training *DispBiNet*, we consider MSE loss between estimated disparities $\hat{D}$ and ground truth D at multiple scales and remove the invalid and occlusion regions with mask map M:

$$\mathcal{L}_{disp} = \sum_{k \in \{L,R\}} \sum_{i=1}^{m} \frac{1}{H_i W_i} ||\hat{D}^k_i - D^k_i||^2 \odot M^k_i, \tag{8}$$

where m is the number of scales of the network and the loss at each scale i is normalized.

4. Stereo Blur Dataset

Currently, there is no dataset specially designed for stereo image deblurring. Therefore, to train our network and verify its effectiveness, we propose a large-scale, multi-scene and depth-varying stereo blur dataset. It consists of a wide variety of scenarios, both indoor and outdoor. The indoor scenarios collect objects and persons, which usually with small depth. The outdoor scenarios include pedestrians, moving traffic and boats as well as natural landscapes. Moreover, we have diversified the dataset by considering various factors including illumination and weather. In the meantime, we have different photograph fashions including handheld shots, fixed shots, and onboard shots, to cover diverse motion patterns.

Inspired by the dynamic scene blur image generation method in [24, 35, 8], we average a sharp high frame rate sequence to generate a blurry image to approximate a long exposure. In practice, we use the ZED stereo camera [1] to capture our data, which has the highest frame rate (60 fps) among the available stereo cameras. However, the frame rate is still not high enough to synthesize look-realistic blur, without generating undesired artifacts which exist in GOPRO dataset [24]. Therefore, we increase the video frame rate to 480 fps using a fast and high-quality frame interpolation method proposed in [25]. Then, we average the varying number (17, 33, 49) of successive frames to generate different blur in size, which is temporally centered on a real-captured sharp frame (ground truth frame). For the synthesis, both two views of the stereo video have the same settings. In addition, to explore how the depth information helps with deblurring, our dataset also provides the corresponding bidirectional disparity of two views, acquired from a ZED camera. We also present the mask map for removing the invalid values in disparity ground truth and occlusion regions obtained by bidirectional consistency check [37].

In total, we collect 135 diverse real-world sequences of dynamic scenes. The dataset consists of 20,637 blurry-sharp stereo image pairs with their corresponding bidirectional disparities at 1280×720 resolution. We divide the dataset into 98 training sequences (17,319 samples) and 37 testing sequences (3,318 samples). The scenarios are totally different for training and testing sets, which avoids the over-fitting problem.

5. Experiments

5.1. Implementation Details

In our experiments, we train the proposed single and stereo image deblurring networks (i.e., *DeblurNet* and *DAVANet*) using our presented Stereo Blur Dataset. For more convincing comparison with single-image methods, we also train and evaluate *DeblurNet* on public GOPRO dataset [24], which contains 3,214 blurry-sharp image pairs (2,103 for training and 1,111 for evaluation).

Data Augmentation. Despite our large dataset, we perform several data augmentation techniques to add diversity into the training data. We perform geometric transformations (randomly cropped to 256×256 patches and randomly flipped vertically) and chromatic transformations (brightness, contrast and saturation are uniformly sampled within $[0.8, 1.2]$) using ColorJitter in PyTorch. To make our network robust, a Gaussian random noise from $\mathcal{N}(0, 0.01)$ is added to the input images. To keep the epipolar constraint of stereo images, we do not adopt any rotation and horizontal flip for data augmentation.

Training. The overall proposed network *DAVANet* contains three sub-networks: *DeblurNet*, *DispBiNet* and *FusionNet*. We first pretrain our *DeblurNet* and *DispBiNet* on each task separately, then add *FusionNet* to the network and train them jointly as a whole. For all models, we set batch size to 2 and use the Adam [15] optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate in our experiments is set to 10^{-4} and decayed by 0.5 every 200k iterations.

For the *DeblurNet*, we first train it on the presented dataset, where 2,000k iterations are sufficient for convergence. For the *DispBiNet*, we first train it using a subset (10,806 samples) of *FlyingThings3D* dataset. In this subset, the samples with large disparity ($> 90\ pixels$) are removed to ensure that the distribution of its disparity is the same as our dataset. Then we finetune the *DispBiNet* fully on our Stereo Blur Dataset until convergence. Finally, we jointly train the overall network on our dataset for 500k iterations.

5.2. Experimental Results

We quantitatively and qualitatively evaluate our single and stereo image deblurring networks (*DeblurNet* and *DAVANet*) on our dataset and compare them with the state-of-the-art deblurring algorithms, including conventional non-uniform deblurring algorithm [39], and CNN-based deblurring methods [36, 6, 24, 16, 44, 38] in terms of PSNR and SSIM. To compare with other end-to-end CNN methods [24, 16, 44, 38], we fully finetune their networks on our dataset until convergence with their released codes. For further comparison, we evaluate our single image deblurring network *DeblurNet* on GOPRO dataset [24] and compare it with aforementioned end-to-end CNN models.

Stereo blur dataset. Although both [24] and [38] propose to use multi-scale recurrent scheme to improve the performance, it inevitably increases the computational cost. To solve this problem, we apply to use two atrous residual blocks and a *Context Module* to obtain the richer feature without a large network in the proposed *DeblurNet*. Table 1 shows that *DeblurNet* outperforms other state-of-the-art single-image deblurring algorithms under the proposed Stereo Blur Dataset. Although the proposed *DeblurNet* performs well with single view, we further evaluate the proposed stereo deblurring network *DAVANet* with other algorithms in Table 1. It demonstrates that the proposed *DAVANet* performs better than the existing dynamic scene methods due to additional depth-aware and view-aggregated features.

Figure 6 shows several examples from the our testing sets. The existing methods [6, 24, 16, 44, 38] cannot perfectly remove the large blur as depth information is not considered in their networks. Although depth information is used in [9], it is hard to estimate it accurately from a single image. In this way, their estimated blur kernels are ineffective and will introduce undesired artifacts into restored images. The proposed *DAVANet* estimates disparity considered as non-uniform prior information to handle spatially variant blur in dynamic scenes. Moreover, it also fuses two-view varying information, which provides more effec-

Table 1: Quantitative evaluation on our Stereo Blur Dataset, in terms of PSNR, SSIM, running time and parameter number. All existing methods are evaluated using their publicly available code. A "-" indicates that the result is not available. Note that the running time for our stereo deblurring network (*DAVANet*) records the forward time of both left and right images.

Method	Whyte [39]	Sun [36]	Gong [6]	Nah [24]	Kupyn [16]	Zhang [44]	Tao [38]	Ours-Single	Ours-Stereo
PSNR	24.84	26.13	26.51	30.35	27.81	30.46	31.65	**31.97**	**33.19**
SSIM	0.8410	0.8830	0.8902	0.9294	0.8895	0.9367	0.9479	**0.9507**	**0.9586**
Time (sec)	700	1200	1500	4.78	0.22	1.40	2.52	**0.13**	**0.31 / pair**
Params (M)	-	7.26	10.29	11.71	11.38	9.22	8.06	**4.59**	8.68

Figure 6: Qualitative evaluations on our Stereo Blur Dataset. The proposed method generates much sharper images with higher PSNR and SSIM values.

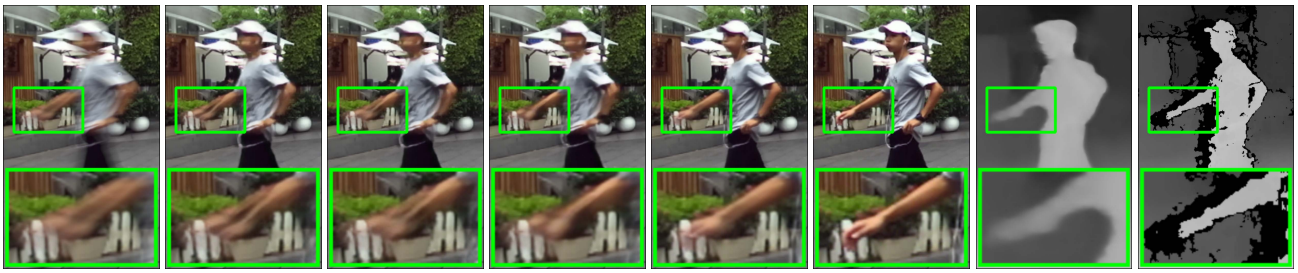
tive and additional information for deblurring. With depth awareness and view aggregation, Figure 6 shows our proposed *DAVANet* can restore sharp and artifact-free images.

GOPRO dataset. Though our single image deblurring network *DeblurNet* performs well on our dataset, we further evaluate it on public GOPRO dataset [24] and compare it with the state-of-the-art CNN models. According to Table 2, the proposed *DeblurNet* with small size outperforms other algorithms in terms of PSNR and SSIM, which further demonstrates the effectiveness of *Context Module*.

Running time and model size. We implement our network using PyTorch platform [32]. To compare running time, we evaluate the proposed method and state-of-the-art image deblurring methods on the same server with an Intel Xeon E5 CPU and an NVIDIA Titan Xp GPU. As traditional blind or non-blind algorithms are used in [39, 36, 6], their methods are time-consuming. With GPU implementation, deep learning-based methods [24, 16, 44, 38] are efficient. To enlarge the receptive field, multi-scale recurrent scheme and large CNN kernel size (e.g. 5×5) are used in [24, 38]. For the same purpose, spatially variant RNNs are used in [44].

Table 2: Quantitative evaluation on the GOPRO dataset [24], in terms of PSNR and SSIM.

Method	Nah [24]	Kupyn [16]	Zhang [44]	Tao [38]	Ours-Single
PSNR	28.49	25.86	29.19	30.26	**30.55**
SSIM	0.9165	0.8359	0.9306	0.9342	**0.9400**



(a) Blurry image (b) Single (c) w/o *DAVA* (d) w/o *VA* (e) Stereo (*DAVA*) (f) Ground Truth (g) Disparity (P) (h) Disparity (T)

Figure 7: The effectiveness of disparity. (a), (f), (g) and (h) denote the blurry image, sharp image, our predicted disparity and ground truth disparity, respectively. (b) and (e) are the deblurring results from the proposed single image deblurring network *DeblurNet* and stereo deblurring network *DAVANet*. In (c), taking two left images as input, *DispBiNet* cannot provide any depth information or disparity for depth awareness and view aggregation. In (d), to only remove the effect of view aggregation, we do not warp the feature from the other view in the *FusionNet*. As the proposed network can estimate accurate disparities and make use of them, it outperforms to the other configurations.

They all lead to long computation time. We find that the proposed *Context Module*, which utilizes convolutions with different dilation rates, can embed multi-scale features and enlarge the receptive field at a low computational cost. In addition, only 3×3 convolutional layers are used in the proposed network which further reduces the size of network. According to Table 1, the proposed network is more efficiency with a small model, compared to the existing CNN-based methods.

5.3. Analysis and Discussions

Effectiveness of the disparity. The proposed model *DAVANet* utilizes estimated disparities in two ways: Depth Awareness (*DA*) and View Aggregation (*VA*). To remove the effect of view aggregation, we do not warp features from the other view in the *FusionNet*, as shown in Figure 7(d). Furthermore, to remove the effect of both depth awareness and view aggregation, we feed two exactly the same images into the proposed network, where no depth information or disparity can be obtained, as shown in Figure 7(c). And we also compare the proposed *DAVANet* with the proposed single image network *DeblurNet*, as shown in Figure 7(b). The Figure 7 demonstrates that the proposed *DAVANet* with depth awareness and view aggregation performs better, using the accurate disparities provided by *DispBiNet*.

Ablation study. The performance improvement of our proposed network should be attributed to three key components, including: *Context Module*, depth awareness, and view aggregation. To demonstrate the effectiveness of each component in the proposed networks, we evaluate the following three variant networks for controlled comparison: (a) To validate the effectiveness of the *Context Module*, we replace the *Context Module* of *DeblurNet* by the one-path convolution block with the same number of layers; (b) To remove the effect of depth information, we remove disparity loss of *DispBiNet* but keep the original input features to *DeblurNet*, where no depth information is involved. The whole network is updated by deblurring losses; (c) To remove the effect of view aggregation, we substitute the concatenation component, the view aggregated features F^L_{views}, with a copy of the reference view features F^L in *FusionNet* (refer to Figure 5 for clarification). We train these networks using the same strategy as aforementioned in Section 5.1. Table 3 shows the proposed network is the best when all components are adopted.

Table 3: Ablation study for the effectiveness of context module, depth awareness and view aggregation. Please see text for details.

Network	w/o Context	**Single**	w/o *DA*	w/o *VA*	**Stereo**
PSNR	31.40	**31.97**	32.69	32.53	**33.19**
SSIM	0.9461	**0.9507**	0.9569	0.9558	**0.9586**

6. Conclusions

In this paper, we present an efficient and effective end-to-end network, *DAVANet*, for stereo image deblurring. The proposed *DAVANet* benefits from depth awareness and view aggregation, where the depth and two-view information are effectively leveraged for spatially-varying blur removal in dynamic scenes. We also construct a large-scale, multi-scene and depth-varying dataset for stereo image deblurring, which consists of 20,637 blurry-sharp stereo image pairs from 135 diverse sequences. The experimental results show that our network outperforms the state-of-the-art methods in terms of accuracy, speed, and model size.

7. Acknowledgements

This work have been supported in part by the National Natural Science Foundation of China (No. 61671182 and 61872421) and Natural Science Foundation of Jiangsu Province (No. BK20180471).

References

[1] Stereolabs. https://www.stereolabs.com/.

[2] Miika Aittala and Frédo Durand. Burst image deblurring using permutation invariant convolutional neural networks. In *ECCV*, 2018.

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, 2018.

[4] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stereoscopic neural style transfer. In *CVPR*, 2018.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2018.

[6] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian D Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi. From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur. In *CVPR*, 2017.

[7] Xinyu Gong, Haozhi Huang, Lin Ma, Fumin Shen, Wei Liu, and Tong Zhang. Neural stereoscopic image style transfer. In *ECCV*, 2018.

[8] Michael Hirsch, Christian J Schuler, Stefan Harmeling, and Bernhard Scholkopf. Fast removal of non-uniform camera shake. In *ICCV*, 2011.

[9] Zhe Hu, Li Xu, and Ming-Hsuan Yang. Joint depth estimation and camera shake removal from single blurry image. In *CVPR*, 2014.

[10] Tae Hyun Kim, Kyoung Mu Lee, Bernhard Scholkopf, and Michael Hirsch. Online video deblurring via dynamic temporal blending network. In *CVPR*, 2017.

[11] Eddy Ilg, Tonmoy Saikia, Margret Keuper, and Thomas Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *ECCV*, 2018.

[12] Daniel S Jeon, Seung-Hwan Baek, Inchang Choi, and Min H Kim. Enhancing the spatial resolution of stereo images using a parallax prior. In *CVPR*, 2018.

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[14] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Schölkopf. Spatio-temporal transformer network for video restoration. In *ECCV*, 2018.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[16] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018.

[17] Dongwoo Lee, Haesol Park, In Kyu Park, and Kyoung Mu Lee. Joint blind motion deblurring and depth estimation of light field. In *ECCV*, 2018.

[18] Bing Li, Chia-Wen Lin, Boxin Shi, Tiejun Huang, Wen Gao, and C-C Jay Kuo. Depth-aware stereo video retargeting. In *CVPR*, 2018.

[19] Lerenhan Li, Jinshan Pan, Wei-Sheng Lai, Changxin Gao, Nong Sang, and Ming-Hsuan Yang. Blind image deblurring via deep discriminative priors. *IJCV*, 2019.

[20] Yunpeng Li, Sing Bing Kang, Neel Joshi, Steve M Seitz, and Daniel P Huttenlocher. Generating sharp panoramas from motion-blurred videos. In *CVPR*, 2010.

[21] Sifei Liu, Jinshan Pan, and Ming-Hsuan Yang. Learning recursive filters for low-level vision via a hybrid neural network. In *ECCV*, 2016.

[22] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

[23] Moritz Menze and Andreas Geige. Object scene flow for autonomous vehiclesk. In *CVPR*, 2015.

[24] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017.

[25] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *ICCV*, 2017.

[26] Mehdi Noroozi, Paramanand Chandramouli, and Paolo Favaro. Motion deblurring in the wild. In *GCPR*, 2017.

[27] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *CVPR*, 2016.

[28] Liyuan Pan, Yuchao Dai, Miaomiao Liu, and Fatih Porikli. Simultaneous stereo video deblurring and scene flow estimation. In *CVPR*, 2017.

[29] Jiahao Pang, Wenxiu Sun, Chengxi Yang, Jimmy Ren, Ruichao Xiao, Jin Zeng, and Liang Lin. Zoom and learn: Generalizing deep stereo matching to novel domains. In *CVPR*, 2018.

[30] Chandramouli Paramanand and Ambasamudram N Rajagopalan. Non-uniform motion deblurring for bilayer scenes. In *CVPR*, 2013.

[31] Haesol Park and Kyoung Mu Lee. Joint estimation of camera pose, depth, deblurring, and super-resolution from a blurred image sequence. In *ICCV*, 2017.

[32] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS Workshops*, 2017.

[33] Anita Sellent, Carsten Rother, and Stefan Roth. Stereo video deblurring. In *ECCV*, 2016.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[35] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *CVPR*, 2017.

[36] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *CVPR*, 2015.

[37] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, 2010.

[38] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, 2018.

[39] Oliver Whyte, Josef Sivic, Andrew Zisserman, and Jean Ponce. Non-uniform deblurring for shaken images. *IJCV*, 2012.

[40] Li Xu and Jiaya Jia. Depth-aware motion deblurring. In *ICCP*. IEEE, 2012.

[41] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *CVPR*, 2013.

[42] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *CVPR*, 2015.

[43] Jiawei Zhang, Jinshan Pan, Wei-Sheng Lai, Rynson WH Lau, and Ming-Hsuan Yang. Learning fully convolutional networks for iterative non-blind deconvolution. In *CVPR*, 2017.

[44] Jiawei Zhang, Jinshan Pan, Jimmy Ren, Yibing Song, Linchao Bao, Rynson WH Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *CVPR*, 2018.

[45] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *ICCV*, 2011.