New End-to-end Network for Stereo High Dynamic Range Imaging

Lifei Zhong and Jiantao Zhou
State Key Laboratory of Internet of Things for Smart City
Department of Computer and Information Science
University of Macau, Macau, China
E-mails: zhonglifei@hotmail.com, jtzhou@um.edu.mo

***Abstract*—Cameras people use in their daily life usually can only obtain low dynamic range (LDR) images. In order to obtain high dynamic range (HDR) images, various methods have been invented. But there is a significant problem with most HDR techniques, namely that original HDR methods require images with different exposure conditions to be taken. In this process, if the captured objects are in motion, the generated HDR image will suffer from ghosting artifacts. To solve this problem, one way is to use different cameras to take images with various exposures simultaneously; by this method the impact of object motion can be minimized. Inspired by this idea, we propose MVMEFNet, an end-to-end network that consists of two sub-networks: Warp Net which is used to align the images taken from two views and produce a disparity map, and Fusion Net which is designed to fuse the aligned left view and right view images. We also innovatively introduce deformable encoder in the Fusion Net, which allows for better error correction of the results in warp net. The experimental results show that our proposed method can obtain stereo HDR image with good visual quality.**

I. Introduction

Nowadays, photography is an important part of our daily life. However, the common commercial cameras can only capture low dynamic range (LDR) image, which causes the image to be far less detailed than the human eye can capture. Therefore many traditional high dynamic range (HDR) technologies were invented to fill this gap[1,2,3]. Most of them need a series of LDR images with different exposures as inputs, and then merge the inputs into one HDR image as output [1]. During this process, if there is only one input device, and the objects in the scene are moving, it will inevitably result in the input LDR images being mismatched, further leading to ghosting in the generated HDR images. A popular idea is to use algorithms to eliminate the impact of this mismatch [2,3]. There is also another idea that simultaneously collects images with different exposures by using additional input devices to reduce the mismatch caused by the object movement [4,5]. This idea not only solves the problems caused by object motion, but also introduces additional benefits; such stereoscopic images or videos generated with this method can be naturally applied to scenarios such as VR/AR. Based on this idea, some researchers have made many attempts. Compared to HDR methods, Stereo HDR(SHDR) methods have extra steps of stereo matching and image warping. Different handcrafted features and cumulative functions [6] have been tried to generate the disparity map. Based on the disparity map, image warping can be applied to the input images to get the warped images, which will be then fed to the appropriate HDR algorithm to produce the final result. Recently deep learning has achieved impressive success in a variety of fields. Several key steps in SHDR are available with deep learning solutions. There have been researchers who proposed deep learning methods of SHDR. Chen et al.[7] used VET-GAN to generate the warped images directly without explicitly generating disparity maps. They then used another HDR fusion GAN to perform the HDR fusion. Although they achieved good results; such an approach has at least two drawbacks. The first one is that this method does not explicitly generate a depth map of the scene, which is certainly a waste. The second one is that this method is not an end-to-end network, which makes it difficult to achieve optimal performance of the framework.

In this paper, in order to exploit the high performance of deep learning while compensating for the shortcomings of previous works, we propose a new end-to-end network for multi-view HDR imaging. Specifically, imitating the steps of the traditional method, our multi-view multi-exposure fusion network(MVMEF-Net) is composed of two sub-networks: Warp-Net and Fusion-Net. In Warp-Net, we perform stereo matching on the initial input images to estimate the disparity map and image warping is then performed based on this disparity map. The results of Warp-Net will be fed into Fusion-Net to generate the final result. The overall network is trained in an end-to-end manner.

Our work has three main contributions: 1) We propose a novel end-to-end learning network for multiple-view multi-exposure image fusion. 2) Our proposed method generates an accurate disparity map, which can be used to generate the depth map of the scene or be used for some other purposes. 3) We propose a component called Deformable Encoder in the Fusion-Net where the Deformable Encoder plays a role of error correction when the inputs have mismatches. This may be an inspiration for other image fusion tasks.

II. Related works

A. Stereo matching

Stereo matching is a classical problem which is related to the depth estimation of stereo images. Many stereo matching methods have been proposed. A traditional stereo matching algorithm [8] needs to compute matching cost, aggregate

matching cost, optimize disparity map and refine final result. Hirschmuller proposed the semi-global matching(SGM) algorithm [6], which has become the standard optimization algorithm for traditional stereo matching methods. Even the early deep learning approaches tried to use CNN features to compute the matching cost and applied SGM to optimize the disparity map. Zbontar and LeCun [9] proposed a method using deep Siamese network to predict the similarity within image patches. Recently, with the development of deep learning, end-to-end approaches have been proposed. Based on the FlowNet [10] proposed by Dosovitskiy et al., Mayer et al. [11] presented an end-to-end DispNet for the estimation of disparity map. Another important improvemen of this task is the introduction of 3D convolution and cost volume into the stereo matching network. Kendall et al. [12] proposed GC-Net, which was the first network for stereo matching using extracted deep features to construct cost volume and using 3D convolutional layer to exploit contextual information. Based on their works, many new approaches have been proposed. Chang et al. [13] proposed PSMNet in which a pyramid pooling module and a stacked hourglass 3D CNN were used to improve the performance. Their method achieved state-of-the-art performance on the KITTI dataset.

B. HDR

The traditional methods for HDR image acquisition could be divided into two main branches: reversing pipeline and multi-exposure fusion. Debevec and Malik [1] tried to reverse the camera pipeline and obtain the radiance of the scene from the RGB image by estimating the camera response function(CRF). After the HDR radiance map is constructed, it is tone mapped into an RGB image that can be shown on common display devices. Merten et al. [14] proposed a technique which can directly assign weights to pixels of the bracketed exposure sequence and fuse input images into a final HDR image without the operation on radiance map. By skipping the camera pipeline reversing, the entire process of HDR image acquisition has been significantly simplified. For static scenes, both of them can achieve good enough results, and the main focus of research in recent years has been on dealing with ghost caused by misalignment for various reasons, such as the movement of scene objects. The popular idea is to accept images with moving objects as input and to eliminate the effect of object motions by algorithms [2,3]. Based on this idea, many deep learning approaches have also been proposed in recent years. Another idea is to use multiple input devices to minimize the object motions in the input image. By establishing a good matching relationship between LDR images of different views, they can be better blended into HDR images without considering the object motions in the input images [4,5]. However, deep learning studies based on this idea have appeared less frequently in recent years. Chen et al. [7] is among the few who have done this with deep learning methods.

III. Proposed Method

This paper proposes a novel end-to-end multi-view multi-exposure network which accepts an LDR stereo image pair as input, as shown in Fig. 1.

A. WarpNet

The WarpNet is designed to accomplish two tasks of stereo matching and image warping, which means that the outputs of this module are accurate disparity maps and warped images. The WarpNet have two parts to fulfill our requirements. The first half of WarpNet is a stereo matching network [13] which can generate the disparity maps we need, while the second half uses the DBI module to generate warped left view image as the input for the next step. The main structure is shown in Fig. 2. The SPP module is used for incorporating global contextual information into image features. These feature maps are later used to construct the cost volume.

The Cost Volume is used to learn matching cost estimation. In this network, it is composed of SPP features by concatenating left and right features according to their correspondence from disparity 0 to the preset maximum value. In this manner we can obtain a 4D Tensor with shape $height \times width \times disparity \times feature\ size$, which is supposed to carry absolute feature representations with semantic information rather than relative representations between features. Given this kind of cost volume, three-dimensional convolutional layers are used to refine this representation due to its ability to learn feature representations from the height, width and disparity dimensions. The last layer of the 3D CNN is a single 3D convolution with a single output channel. This cost volume is then upsampled to obtain a final cost volume with size $H \times W \times D$. After obtaining such a cost volume, disparity regression is applied to generate a dense disparity map. For each disparity d, the probability volume is converted from taking the negative of predicted costs c_d and normalized with

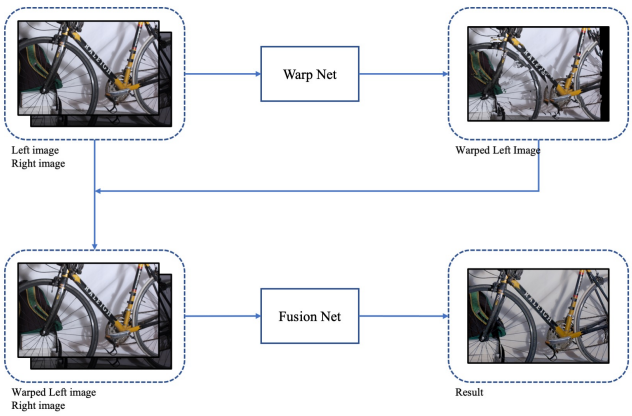


Fig. 1. The overall structure of the proposed MVMEFNet. The MVMEFNet consist of 2 sub-network. WarpNet accepts a pair of left view and right view images, which are then processed into a pair of warped left view images and right view images. The processed image pairs are further fed into FusionNet for estimating the final HDR image.

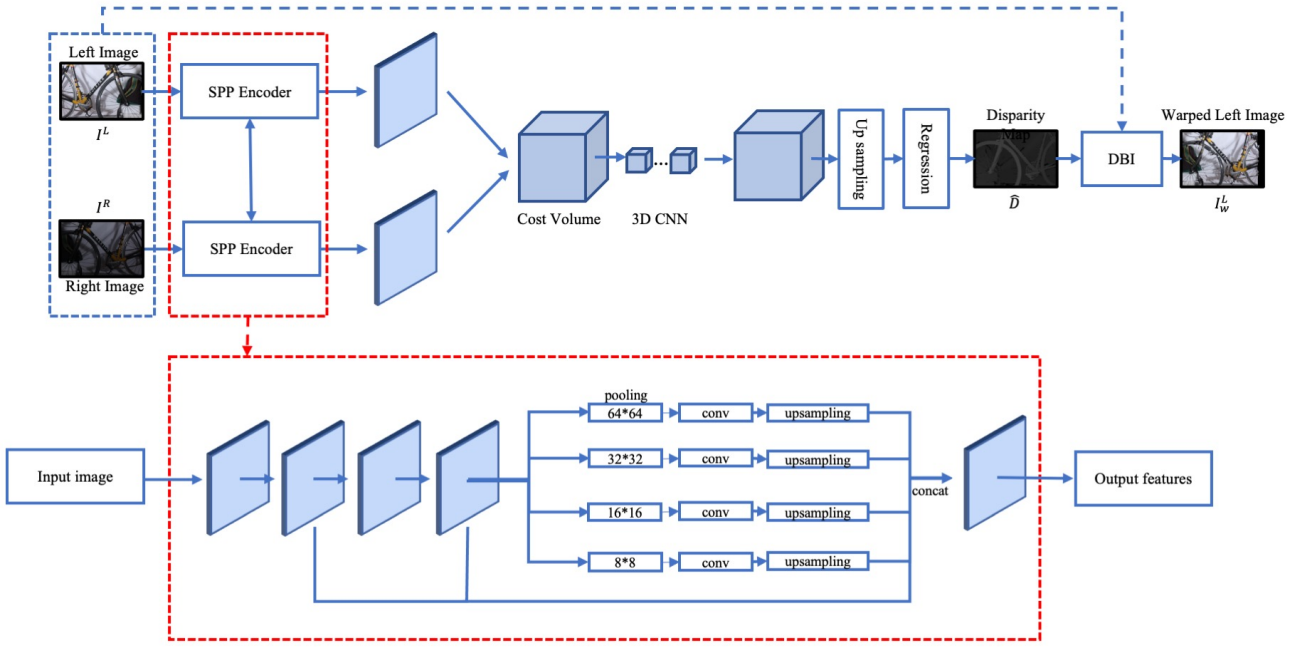


Fig. 2. The detailed structure of the WarpNet. The SPP module is a feature extraction module which can incorporate features in different level and has been proved to be effective in task of stereo matching. Note that constructing cost volume, passing through a 3D CNN, performing a disparity regression is a standard process for stereo matching which. DBI is an abbreviation for the Differentiable Bicubic Interpolation.

the softmax operation, $\sigma(\cdot)$. Then the predicted disparity $\hat{d}$ is defined by soft argmin, as

$$\hat{d} = \sum_{d=0}^{D_{max}} d \times \sigma(-c_d) \tag{1}$$

This function is fully differentiable. It is proved in [13] that using this function allows us to predict disparity in a regression way and can improve performance.

Assembling disparity $\hat{d}$ in all positions, we get the disparity map $\hat{D}$ with resolution of $H \times W$. To make our final network to be end-to-end trainable, the differentiable bicubic interpolation [16] is employed to implement the process of warping. Using the disparity map $\hat{D}$ and the left view image I^L as inputs of the DBI module, the warped left view image I_w^L can be finally obtained.

B. FusionNet

In the previous sub-network, we have obtained the warped left view image I_w^L which is supposed to be well aligned with the right view image I^R. The purpose of our FusionNet is to accept a pair of aligned images I_w^L and I^R as input and then output final HDR images $\hat{H}$. In order to achieve this goal well, our FusionNet consists of two parts: the attention module and the merging module [17]. The structure of this network is shown in Fig. 3. When an aligned image pair is provided as input, it first goes through a deformable convolutional layer to extract features, as

$$Z_1 = DConv(I_w^L), Z_2 = DConv(I^R) \tag{2}$$

Both Z_1 and Z_2 have 64 channels. The Deformable Convolutional Network(DCN) was first proposed in [18]. It was proved that by replacing the traditional CNN with DCN, the object detection and semantic segmentation tasks can get a performance improvement. It is also easy to replace the 2d convolutional layer with deformable convolutional layer because they have the same input and output. The experiments in ablation studies will further show the effectiveness of the deformable convolutional layer in our task. After feature maps Z_1 and Z_1 are extracted, they are concatenated and then passed into a two-layer convolutional neural network $Attention$ with a final sigmoid activation:

$$W = \sigma(Conv(Concat(Z_1, Z_2))) \tag{3}$$

As a result, an attention map W in the range between 0 and 1 can be obtained. The attention maps have the same channels with the input feature maps and then are point-wise multiplied with left feature map:

$$Z_1^{'} = W \odot Z_1 \tag{4}$$

The attention-guided left feature maps $Z_1^{'}$ and right feature maps Z_2 are then concatenated to a new feature maps Z_m and passed to the merging module:

$$Z_m = Concat(Z_1^{'}, Z_2) \tag{5}$$

The dilated residual dense block(DRDB) was proposed in [17] and proved to be effective in HDR imaging. In this work, 3 sequential DRDBs are used, and each DRDB consists of 6 2-dilated convolutions[19]. Before reconstructing the final HDR

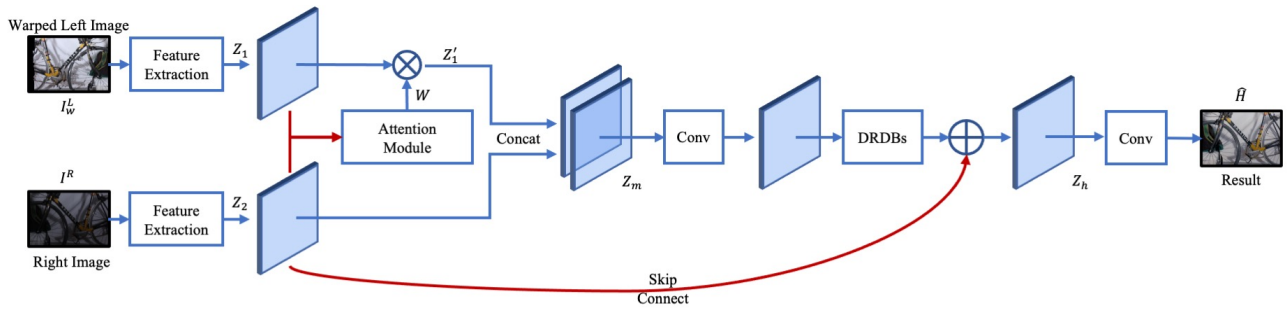


Fig. 3. The detailed structure of the FusionNet. Here we use a single deformable convolutional layer to extract features. The effectiveness of deformable convolutional layer is shown in our experiments. The attention module is exlude the regions with poor quality in warped left view image. The concatenated features is passed through 3 dialated residual blocks(DRDB) and then summed with skip connected right view features. The final HDR image is obtained by 2 layer convolution without any post-processing .

image, a global skip connect is used to make the merging module tend to learn the residual features. This strategy is inspired by the super-resolution tasks[20] and its effectiveness in HDR imaging has also been studied in [17]. The final feature maps Z_h is then passed through two convolutional layers with activations. The final HDR image $\hat{H}$ is estimated without any extra post-processing or tone mapping.

$$\hat{H} = Conv(Z_h) \tag{6}$$

C. Loss function

As stated above, our proposed MVMEFNet estimates disparity map $\hat{D}$ and final HDR image $\hat{H}$ as our outputs. Our network is trained by minimizing ℓ_1 distance between the outputs and the ground-truth disparity and ground-truth HDR image respectively:

$$\ell = \|\hat{D} - D_{gt}\|_1 + \|\hat{H} - H_{gt}\|_1 \tag{7}$$

where D_{gt} is the ground-truth disparity map and the H_{gt} is the ground-truth HDR image.

IV. Experiments

A. Setups and training details

As our work requires inputs of stereo images with different exposures, we use Middlebury 2014 stereo Dataset[21] to construct our training and testing datasets. The official Middlebury dataset only offers disparity ground-truth, therefore we use a recently proposed static MEF method[22] to generate the ground-truth of our final result. The Middlebury dataset provides about 22 scenes and there are about 4 different lighting conditions for each scene, and 3-8 exposure settings for each lighting condition. We choose 18 scenes for training and 3 scenes for testing. As a result, there are about 50 training image sets and 10 testing image sets. For each image set, we manually choose the input pairs and generate the corresponding ground-truth with them. All chosen input image pairs follow the pattern that the left view image is over-exposed and the right view image is under-exposed. Since the original maximum disparity is outside of our capacity, we first downsamples the original image to make its height and width half. We randomly crop them into patches with the size of 256*512 as the input of our network. In the training stage, we use the Adam optimizer and the learning rate is set as 1e-3. In the testing stage, the PSNR and SSIM values are computed as reference for quantitative analysis and comparison.

B. Comparisons with SOTA

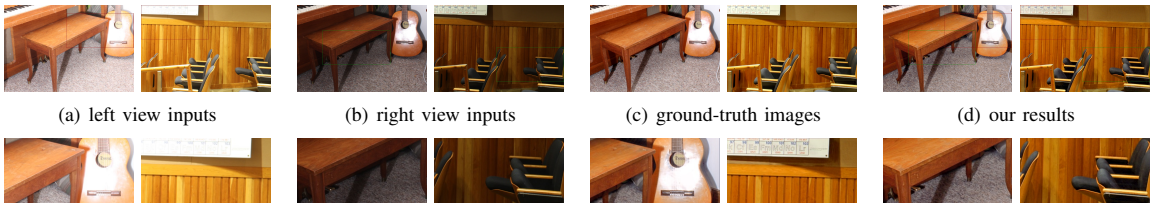
Since the VET-GAN [7] is the only end-to-end deep learning multi-view HDR method and there is no comparison with other methods in their paper, we only compares the proposed method with VET-GAN. They uses the same dataset as our work. Table 1 shows the detailed quantitative results. On the same dataset, our method achieves a PSNR improvement of 0.41 and an SSIM improvement of 0.03. In addition, as mentioned before, our method can generate accurate dense disparity maps. Here we show an example of our results in Fig. 4. Our method has good performance in both the over-exposed(OE) areas such as the bottom of the chair and under-exposed(UE) areas such as the surface of the guitar.

C. Ablation Studies

To demonstrate the effectiveness of the Deformable convolutional layer, we compare the average PSNR/SSIM over the test set under the same training conditions in Table I. We can see that the average PSNR values increase about 2.0544 and the average SSIM increase about 0.0167 by replacing the 2D convolutional encoder with deformable convolutional encoder. In Table II and Table III, we show detailed quantitative results on each of the scenario in the test set. In addition, visual results of these 2 methods are shown in Fig. 5. The overall color of results with DCN is closer to that of the original image. The results with DCN have less ghosting artifacts, such as the area in the upper left corner.

TABLE I
QUANTITATIVE COMPARISON

Method	PSNR	SSIM
VET-GAN	28.8665	0.91
Ours	29.2798	0.91
Ours w/o DCN	27.2254	0.89



(a) left view inputs (b) right view inputs (c) ground-truth images (d) our results

(e) left view over-exposed areas (f) right view under-exposed areas (g) over-exposed areas of results (h) under-exposed areas of results

Fig. 4. Visual presentation of samples of our results

(a) inputs

(b) ground truth

(c) result w/o DCN (27.46) (d) result w DCN (29.91)

Fig. 5. Comparison of different convolutional layer.

V. Conclusions

In this paper, a novel end to end learning method for multi-view multi-exposure fusion is proposed. The MVMEF-Net consists of 2 sub-network. The first sub-network could generate a disparity map and a warped left view image. The second sub-network accept warped left view image and right view image as input and then generate the final output. By combining these 2 sub-network, the whole network can not only be trained with end-to-end manner, but also outperform other methods of these 2 individual tasks.

TABLE II
QUANTITATIVE COMPARISON(PSNR/SSIM) ON EACH SCENARIO

Method	Bicycle	Classroom	Piano
MVMEFNet	28.79/0.91	29.96/0.89	28.96/0.88
MVMEFNet w/o DCN	27.40/0.91	26.79/0.92	27.53/0.90

TABLE III
QUANTITATIVE COMPARISON(PSNR/SSIM) OVER DIFFERENT LIGHT CONDITIONS ON CLASSROOM

Method	L1	L2	L3	L4
MVMEFNet	32.64/0.92	28.89/0.90	28.52/0.93	29.80/0.93
MVMEFNet w/o DCN	30.84/0.93	26.34/0.88	26.69/0.85	23.31/0.91

Acknowledgment

This work was supported by Macau Science and Technology Development Fund under 077/2018/A2, by Research Committee at University of Macau under MYRG2018-00029-FST and MYRG2019-00023-FST, by Natural Science Foundation of China under 61971476.

References

[1] P. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, 1997, pp. 369–378.
[2] T. Grosch, "Fast and robust high dynamic range image generation with camera and object movement," *Vision, Modeling and Visualization, RWTH Aachen*, vol. 277284, 2009.
[3] K. Jacobs, C. Loscos and G. Ward, "Automatic high-dynamic range image generation for dynamic scenes," In *IEEE Computer Graphics and Applications*, vol. 28, no. 2, pp. 84-93, 2008.
[4] H. Lin and W. Chang, "High dynamic range imaging for stereoscopic scene representation," In *IEEE International Conference on Image Processing*, 2009, pp. 4305-4308.
[5] W. Park *et al.*, "Stereo vision-based high dynamic range imaging using differently-exposed image pair," *Sensors*, vol. 17, no. 7, pp. 1473, 2017.
[6] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 807-814, vol. 2, 2005.
[7] Y. Chen *et al.*, "New stereo high dynamic range imaging method using generative adversarial networks," In *IEEE International Conference on Image Processing*, 2019, pp. 3502-3506.
[8] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1, pp. 7-42, 2002.
[9] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp.2287-2318, 2016.
[10] A. Dosovitskiy *et al.*, "FlowNet: learning optical flow with convolutional networks," In *IEEE International Conference on Computer Vision*, 2015, pp. 2758-2766.
[11] N. Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4040-4048.

[12] A. Kendall *et al.*, "End-to-end learning of geometry and context for deep stereo regression," In *IEEE International Conference on Computer Vision*, 2017, pp. 66-75.
[13] J. Chang and Y. Chen, "Pyramid stereo matching network," In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410-5418.
[14] T. Mertens, J. Kautz and F. Van Reeth, "Exposure fusion," In *Pacific Conference on Computer Graphics and Applications*, 2007, pp. 382-390.
[15] K. He, X. Zhang, S. Ren and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904-1916, 2015.
[16] M. Jaderberg, K. Simonyan, A. Zisserman, "Spatial transformer networks." In *Advances in Neural Information Processing Systems*, 2015.
[17] Q. Yan *et al.*, "Attention-guided network for ghost-free high dynamic range imaging," In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1751-1760.
[18] J. Dai *et al.*, "Deformable convolutional networks," In *2017 IEEE International Conference on Computer Vision*, 2017, pp. 764-773.
[19] F. Yu, and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.
[20] Y. Zhang, Y. Tian, Y. Kong, B. Zhong and Y. Fu, "Residual dense network for image super-resolution," In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472-2481.
[21] D. Scharstein *et al.*, "High-resolution stereo datasets with subpixel-accurate ground truth." In *German conference on pattern recognition*, Springer, Cham, 2014.
[22] H. Li, K. Ma, H. Yong and L. Zhang, "Fast multi-scale structural patch decomposition for multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 5805-5816, 2020.