ADNet: Attention-guided Deformable Convolutional Network for High Dynamic Range Imaging

Zhen Liu[1,2] Wenjie Lin[1] Xinpeng Li[1] Qing Rao[1] Ting Jiang[1]
Mingyan Han[1] Haoqiang Fan[1] Jian Sun[1] Shuaicheng Liu[3,1*]
[1]Megvii Technology [2]Sichuan University
[3]University of Electronic Science and Technology of China
https://github.com/Pea-Shooter/ADNet

Abstract

In this paper, we present an attention-guided deformable convolutional network for hand-held multi-frame high dynamic range (HDR) imaging, namely ADNet. This problem comprises two intractable challenges of how to handle saturation and noise properly and how to tackle misalignments caused by object motion or camera jittering. To address the former, we adopt a spatial attention module to adaptively select the most appropriate regions of various exposure low dynamic range (LDR) images for fusion. For the latter one, we propose to align the gamma-corrected images in the feature-level with a Pyramid, Cascading and Deformable (PCD) alignment module. The proposed ADNet shows state-of-the-art performance compared with previous methods, achieving a PSNR-l of 39.4471 and a PSNR-μ of 37.6359 in NTIRE 2021 Multi-Frame HDR Challenge.

1. Introduction

High dynamic range imaging technique aims at recovering an HDR image from one or several LDR images. The former refers to single-frame HDR imaging [6, 18, 20, 28], while the latter refers to multi-frame HDR imaging [14, 22, 23, 32, 33]. It has drawn much attention from low-level vision communities as traditional photography sensors cannot capture the actual dynamic range in nature scenes [24, 30]. Compared with single-frame HDR imaging, multi-frame HDR imaging is more practical and promising due to its informative bracket LDR inputs. If the LDR images are aligned perfectly, i.e., no object motion and camera jittering, the static images can be well fused [22, 23]. When photographing with hand-held cameras, we need to handle the misalignments of various exposure LDR images first apart from image fusion.

*Corresponding author.

(a) (b)

Figure 1. (a) LDR inputs with varying exposures. (b) Our result compared with the previous representative method AHDRNet. We show differences in the zoomed-in patches. Our result is free from noise and ghost artifacts while hallucinates more accurate details over the saturated areas. Zoom-in for a better comparison.

Existing explicit image alignment methods mainly consists of three categories: global alignment with homography [10], middle-level alignment with meshflow [19], and pixel-level alignment with optical flow [1]. However, homography and meshflow cannot align foreground dynamic objects, and optical flow is erroneous in the presence of occlusion [21]. Several traditional methods are proposed to detect misaligned regions and then reject these pixels as outliers during the fusion process. However, these methods are often prone to introducing ghost artifacts as accurately identifying the dynamic objects is difficult.

Recently, several learning-based methods have been explored. Kalantari *et al.* proposed the first deep convolu-

tional neural network (CNN) for HDR imaging of dynamic scenes. They first aligned the LDR images with optical flow and then fused the aligned images by a CNN [14]. However, optical flow is unreliable when occlusion and saturation occurs as mentioned above. Wu *et al.* proposed the first non-flow-based approach, which performed homography alignment on LDR images before fed them into a CNN [32]. Yan *et al.* offered to handle motion by an attention module which achieved state-of-the-art results [33] and later introduced a non-local neural network [35]. All these methods processed the input LDR images and their gamma-corrected images uniformly, i.e., by simply concatenating them, resulting in blurry or ghost artifacts frequently.

In this paper, we propose an Attention-guided Deformable Convolutional Network (ADNet), a new pipeline to tackle such problems. Instead of directly concatenating the LDR images and their gamma-corrected images, we propose to process them with dual branches. Specifically, a spatial attention module is used for extracting the LDR images' attention features for better fusion, and a Pyramid, Cascading and Deformable (PCD) alignment module [31] is adopted to align the gamma-corrected images in the feature-level. Such design is motivated by the intuition that the images in the LDR domain help detect the noisy or saturated regions while the HDR counterparts help to detect misalignments [14].

Fig. 1 illustrates the results of our method and the state-of-the-art method AHDRNet [33]. As can be seen, the results of AHDRNet fail to recover the details of saturated regions. In contrast, our method produces noise-free and ghost-free results while hallucinating more accurate contents in over saturated areas. Our main contribution can be summarized as follow:

- We propose a novel dual-branch pipeline for multi-frame HDR imaging of dynamic scenes. Unlike previous methods that treat the LDR and gamma-corrected images uniformly, we process the LDR images with a spatial attention module and process the corresponding gamma-corrected images with a PCD alignment module.
- Existing learning-based methods use either optical flow based aligment or no explicit alignment. In this paper, we propose to align the dynamic frames with deformable alignment module, showing significant improvements over previous opponents. To our best knowledge, this is the first application of deformable convolutions for multi-frame HDR imaging.
- Experimental results show that the proposed method achieves better results than the state-of-the-art methods, both quantitatively and qualitatively. Our approach also achieves the best results in NITRE 2021 Multi-Frame HDR Challenge.

2. Related Works

We briefly summarize existing approaches into three categories: motion rejection methods, image registration methods, and learning-based methods.

Motion rejection methods perform global alignment upon LDR images first and then reject misaligned pixels before the image fusion. Gallo *et al.* proposed to predict colors in various exposures directly and then compared them with original values to detect motion [8]. Grosch *et al.* calculated an error map according to the differences of alignment colors to reject misaligned pixels [9]. Pece *et al.* detected the motion regions by computing the median threshold bitmap for input LDR images [26]. Jacobs *et al.* detected the misalignment regions by weighted intensity variance measurement [12]. Zhang *et al.* computed a weight map of the LDR inputs in the gradient domain [37]. Khan *et al.* proposed to calculate the probability maps for pixels that belong to the background [16]. Oh *et al.* also introduced rank minimization to detect ghost regions [25]. These methods often lead to unsatisfactory HDR effects as rejecting pixels will drop helpful information.

Image registration approaches register local image regions for fusion. Bogoni *et al.* proposed to predict motion vectors using optical flow [2]. Hu *et al.* proposed to perform image alignment in the transformation domain based on brightness and gradient consistencies [11]. Kang *et al.* converted LDR image intensities to the luminance domain according to exposure time and then calculated optical flow to compensate motion [15]. Zimmer *et al.* first registered the LDR images with optical flow and then recovered the HDR image [38]. Patch-based optimization is another type of image registration method besides optical flow. Sen *et al.* proposed to align LDR images and reconstruct HDR image in a joint energy optimization process [29]. Jinno *et al.* proposed to model the displacement with Markov random field [13]. Image registration methods show better performance than motion rejection approaches. However, if large motion occurs, this approach introduces apparent artifacts.

Several deep learning approaches have been proposed recently [6,7,14,32–34]. Single frame based methods recover HDR from a single LDR image. Eilertsen *et al.* directly predicted HDR image from a single LDR input through a deep CNN [6]. Endo *et al.* generated bracket LDR images from a single frame first and then fuse them to reconstruct HDR image [7]. Kalantari *et al.* proposed the first deep multi-frame HDR imaging method of dynamic scenes. The LDR images were first aligned with optical flow and then blended by a fusion subnet [14]. Wu *et al.* proposed the first non-flow-based deep framework. They performed a global registration on LDR inputs using homography and handled alignments and fusion through a UNet-based network [32]. Yan *et al.* proposed an attention-based deep CNN to control large motion, achieving the state-of-the-art

eprformance [33]. Our proposed method is built upon deep neural networks.

3. Method

The problem of multi-frame HDR imaging is to reconstruct an HDR image from several LDR images with various exposures. The middle frame of the LDR images is commonly selected as the reference image for motion alignment. In this paper, we consider 3 LDR images, i.e., $I_i, i = 1, 2, 3$, as input and let the second LDR image I_2 be the reference image I_r. Existing learning-based methods [14, 32, 33, 35] first map the input LDR images to the HDR domain using gamma correction and then concatenate them directly as the network input:

$$\check{I}_i = \frac{(I_i)^\gamma}{t_i}, \quad i = 1, 2, 3 \tag{1}$$

where t_i is the exposure time of I_i, γ is the gamma correction parameter, and $\check{I}_i$ denotes the corresponding gamma-corrected images.

Instead of concatenating them directly, we propose a novel dual-branch pipeline. Specifically, for the LDR images, we extract the attention feature maps with a spatial attention module $\mathcal{A}$. As for the gamma-corrected images, we adopt a PCD align module $\mathcal{P}$ to handle dynamic objects or camera motions. Therefore, our proposed network f can be defined as:

$$I^{\mathrm{H}} = f(\mathcal{A}(I_i), \mathcal{P}(\check{I}_i); \theta) \tag{2}$$

where I^H denotes the reconstructed HDR image and θ denotes the network parameters.

3.1. Network Structure

We present the network structure of the proposed ADNet in Fig. 2. The overall structure of ADNet mainly consists of three components: spatial attention module for LDR images (Fig. 2 (a)), PCD align module for gamma-corrected images (Fig. 2 (b)) and the fusion subnet for HDR reconstruction (Fig. 2 (c)). We first utilize a spatial attention module to extract the attention feature maps of three LDR images. Meanwhile, we use a PCD align module to align the gamma-corrected images in the feature level. Finally, we concatenate the attention features and the aligned gamma-corrected features in order of exposure time and feed them to the fusion subnet, generating the estimated HDR image.

Spatial Attention Module for LDR Images. Given the LDR images $I_i, i = 1, 2, 3$ with the shape of HxWx3, we first extract their LDR features by a single convolutional layer. For each non-reference LDR image $I_i, i \neq 2$, we concatenate the LDR feature with the feature of reference image as the input of the spatial attention module, generating the attention map with the range of 0-1. We then compute the element-wise multiplication of the LDR feature and its corresponding attention map to generate the spatial attention feature of each LDR image. The process can be formulated as

$$f_i^t = \mathcal{A}(I_i, I_r), \quad i = 1, 3 \tag{3}$$

where f_2^t is the reference feature. In this paper, we adopt the attention module as used in [33]. The details are shown in Table 1.

Table 1. Details of the attention module used in our NTIRE 2021 HDR Challenge (Multi-Frame Track) submission.

Name	# Out	Type
Input	128	Input Feature Maps
Conv-Layer-1	128	Conv 3x3
ReLU-Layer-2	128	ReLU Activation
Conv-Layer-3	64	Conv 3x3
Sigmoid-Layer-4	64	Sigmoid Activation

PCD Align Module for Gamma-corrected Images. Performing alignment at the feature level is better than at the image level [3]. We also adopt this by employing deformable alignment [4]. Specifically, after applying gamma correction on the LDR images, we first extract the pyramid features using Stride convolutions and then perform deformable alignment with the reference feature on each scale of features, which is the same as the PCD module proposed by [31], i.e.,

$$f_i^l = \mathcal{P}(\check{I}_i, \check{I}_r), \quad i = 1, 3 \tag{4}$$

Fusion Subnet for HDR Reconstruction. The feature maps f_i^t and f_i^l are concatenated together as the input of the fusion subnet. Here, f_2^t and f_2^l denote the same reference features, which are not processed by the alignment or attention modules. The fusion subnet consists of several dilated residual dense blocks (DRDBs) which are also used in [33]. The usage of dilated convolution [36] increases the receptive field. We also use local skip connection and global skip connection for better training our model. The aforementioned network design generates ghost-free and noise-free HDR results, together with clearer image contents.

3.2. Training Strategy

Gamma Disturbance Existing methods usually first linearize the non-linear LDR images using the camera response function (CRF) [5] and then apply gamma correction(e.g., $\gamma = 2.2$) on these linearized images to produce the input images [14, 29]. However, as the CRF of each camera is not strictly the same, a fixed gamma value may

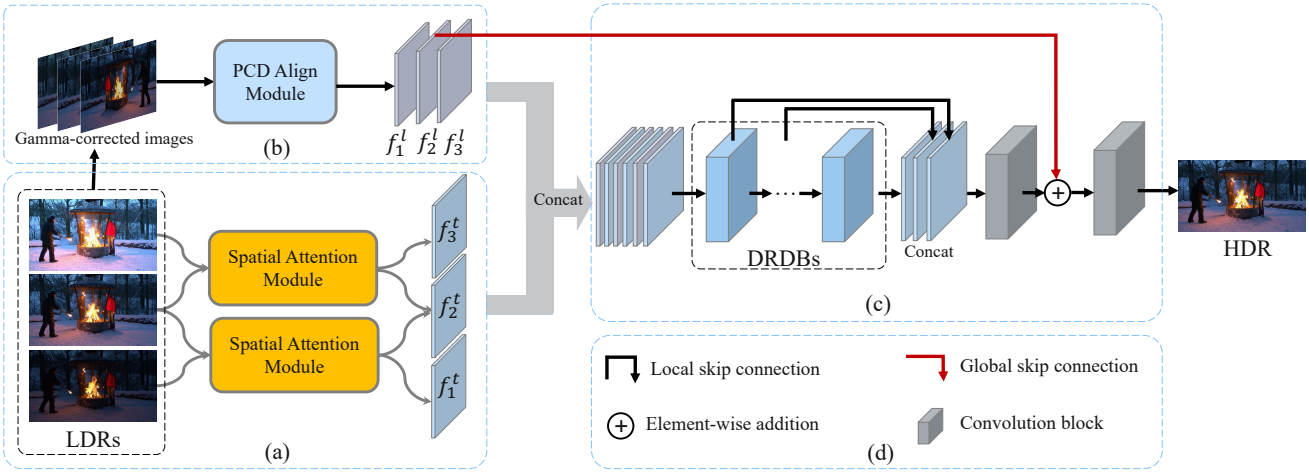


Figure 2. The pipeline of our method. The network mainly consists of three components: (a) The input LDR images are first fed into a spatial attention module to generate attention feature maps, and then (b) we adopt a PCD module to align the corresponding gamma-corrected images, (c) finally we employee several dilated residual dense blocks in the fusion subnet.

not always be the most appropriate. In this paper, we propose a gamma disturbance strategy. Specifically, instead of maintaining a fixed gamma value all the time, we randomly apply a gamma value of 2.24 ± 0.1 with the probability of 30%. By adopting this strategy, The proposed method obtains about 0.1dB gain in terms of PSNR.

Loss Function For multi-frame HDR imaging tasks, optimizing the network on the tonemapping domain is more effective than optimizing directly in the HDR domain as the HDR images are usually viewed after tonemapped [14]. We also adopt such a strategy to train our ADNet. Given an estimated HDR image I^H and the corresponding ground-truth HDR image I^{GT}, we first apply tonemapping onto them using the commonly used μ-law:

$$\mathcal{T}(x) = \frac{\log(1+\mu x)}{1+\mu}. \tag{5}$$

We set $\mu = 5000$ in this paper. Then we compute the l_1 error of the tonemapped images as our loss function, i.e.,

$$\mathcal{L} = \left\| \mathcal{T}\left(I^{\mathrm{GT}}\right) - \mathcal{T}\left(I^{H}\right) \right\|_1 \tag{6}$$

4. Experiments

4.1. Dataset and Implementation Details

We train and evaluate our method on the dataset provided by the NTIRE2021 Multi-Frame HDR Challenge [27]. It contains 1463 valid scenes in total as we exclude 31 incomplete scenes. We select 100 scenes as the validation set and keep the remaining as the training set. Each scene consists of three LDR images with various exposures and their corresponding HDR ground truth.

During the training stage, we first crop the input LDR images to 256x256-sized patches with a stride of 128. The network is optimized by an Adam optimizer [17] with initial learning rate of 1e-4 and decay rate of 0.1, we set $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. Our experiments are implemented in PyTorch and trained on 8 NVIDIA 2080Ti GPUs with batch size of 16. We train the model from scratch for 200 epochs, and the whole training costs about three days. We select the best model using the PSNR-μ score calculated on our validation set when the training reaches plateaus.

The entire testing process is conducted on a single 2080Ti GPU. Constrained by the limited GPU memory, we split each test image into size 1060x1000 and 1060x900 for testing and then concatenate them to the full size 1060x1900. We compute the PSNR-l and PSNR-μ scores as the testing metrics where '-l' and '-μ' means the ones computed in the linear domain and the tonemapped domain, respectively.

Table 2. Quantitive comparison with AHDRNet. The PSNR-l and PSNR-μ refer to the PSNR scores computed in the linear domain and tonemapped domain. The 'TTA' means testing-time augmentation.

		PSNR-l	PSNR-μ
AHDRNet	w/o. TTA	38.6737	36.7068
	w/ TTA	39.0577	36.8073
Ours	w/o. TTA	39.3398	37.2068
	w/ TTA	**39.9167**	**37.3548**

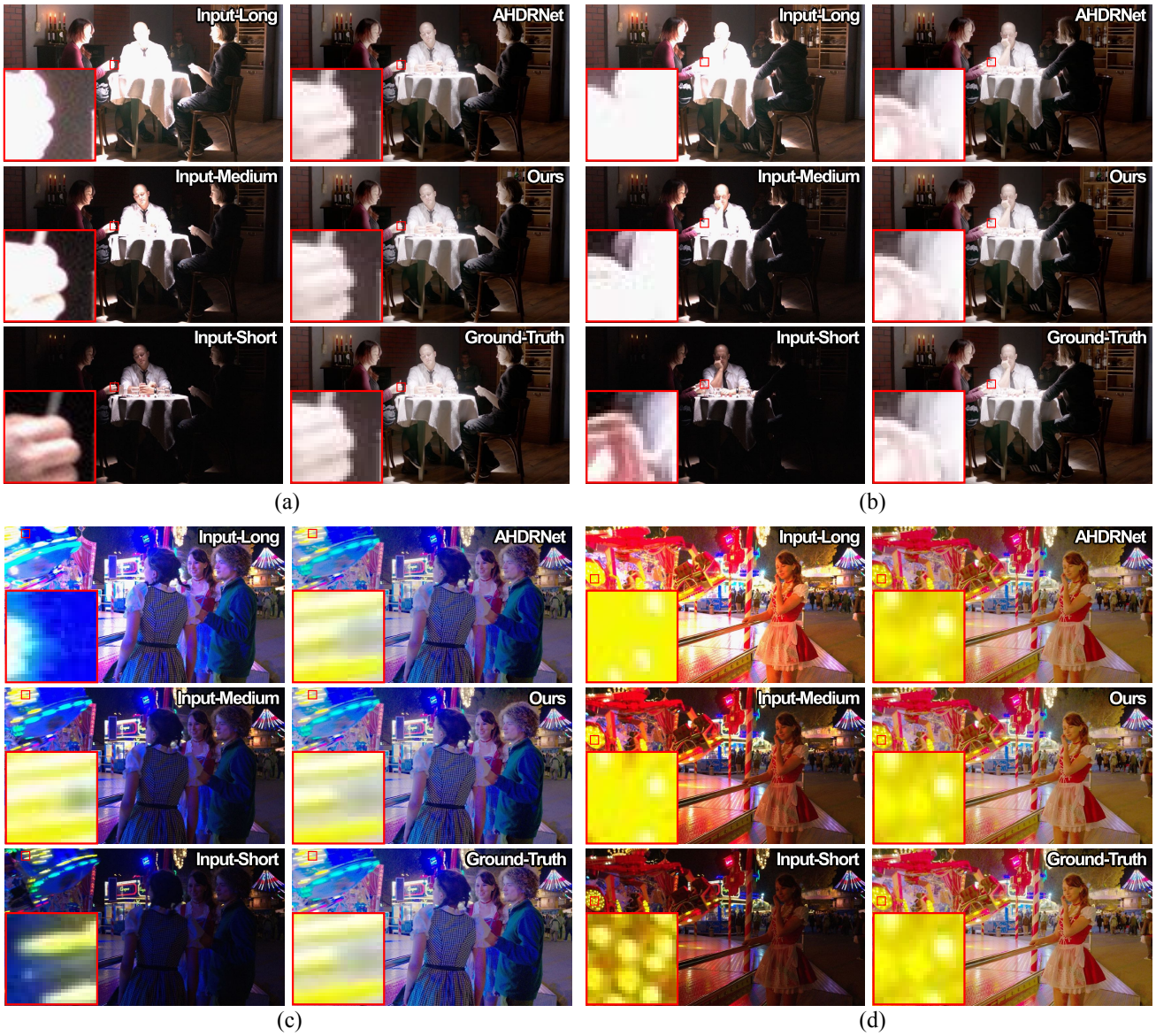
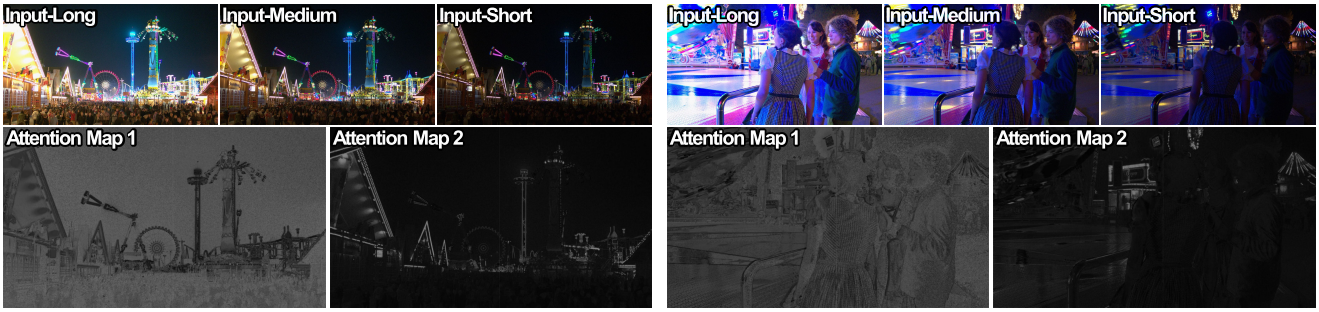


Figure 3. Qualitative Comparison with existing state-of-the-art method AHDRNet. The proposed ADNet can not only reconstruct clearer image contents in motion boundaries (Fig. 3 (a) and Fig. 3 (b)) but also hallucinate more reasonable details in saturated regions (Fig. 3 (c) and Fig. 3 (d)).

4.2. Results and Analysis

To demonstrate the superiority of our proposed ADNet, we compare it with the existing state-of-the-art method AHDRNet [33], both quantitively and qualitatively. For fair comparisons, we retrain AHDRNet in the challenge data with the same settings as ours and report the PSNR computed in the linear domain and tonemapped domain, i.e., PSNR-l and PSNR-μ. As listed in Table 2, the proposed ADNet produces higher scores among all the calculated metrics, outperforming the AHDRNet by 0.85dB in terms of PSNR-l and 0.55dB in terms of PSNR-μ. The 'TTA' means testing-time augmentation, which can be seen as a self-ensemble strategy. In this paper, we apply a '4x' version where the final result is averaged from four versions of the input images, i.e., the original, the permuted, the vertically flipped, and the horizontally flipped images.

We also report the qualitative results compared with AHDRNet. As illustrated in Fig. 3, The first two scenes (Fig. 3 (a) and Fig. 3 (b)) contain subtle motion upon the lady's hands, and meanwhile, the middle frame and the long-time exposure frame encounter saturation in these regions. The



(a) (b)

Figure 4. Visualization of attention maps. The attention maps adaptively select the most appropriate regions of LDR images to fusion. In each scene of Fig. 4 (a) and Fig. 4 (b), the first attention map tends to suppress the over-exposure regions and vice-versa of the second one. Zoom-in for a better view.

result of AHDRNet produces undesirable details and blurry contents while ours are clearer and more precise in motion boundaries. Fig. 3 (c) is more intractable as the motion magnitude gets larger upon the saturated hobbyhorse. The scene of Fig. 3 (d) encounters more severe saturation regions. As seen, the results of AHDRNet fail to recover the corresponding areas while our method can hallucinate reasonable details and is free of ghost artifacts and noises.

We attribute the de-ghosting and HDR reconstruction ability of our method to the new network design. On the one hand, the spatial attention module performed on the LDR images can adaptively select proper image regions for fusion, i.e., the light regions in the short-time exposure frame and the dark region in the long-time exposure frame. To futher verify this, we visualize the attention maps by averaging them along the channel dimension. As shown in Fig. 4, the attention maps suppress the over-/under-exposure regions and focus on the well-exposed areas. On the other hand, the PCD align module handles the gamma-corrected images which contain camera motions or dynamic objects and aligns them more accurately. We also verify the effective network design through ablation studies.

4.3. Ablation Study

To verify the effectiveness of the proposed ADNet, we conduct ablation studies of the network architecture and analyze the results. It should be noted that all the ablation studies are compared in our validation set as the test set is currently unavailable. We compare our method with three variants as follow:

- **Baseline**. Our method takes the AHDRNet [33] as our baseline, which contains an attention network and a merging network.
- **Variant 1**. This variant replaces the attention module in the baseline as a vanilla deformable alignment module. Specifically, we apply deformable alignment only on the original scale features instead of pyramid features.
- **Variant 2**. This variant adopts a PCD alignment module instead of a single scale one as used in the first variant.
- **Ours**. The entire network architecture of the proposed dual-branch ADNet, which contains a spatial attention module and a PCD align module.

Table 3. Ablation studies on the network architecture.

	PSNR-l	PSNR-μ
Baseline	41.7593	34.6083
Variant 1	42.0612	34.6113
Variant 2	43.2681	34.7316
Ours	**43.6218**	**34.8606**

PCD Align V.S. Vanilla Deformable Align As shown in Table 3, the first variant only employs the vanilla deformable alignment with one single scale while the second variant adopts a PCD align module as used in [31]. Compared with the baseline model, the deformable alignment shows better performance. With the usage of pyramid features alignment, the second variant has a 1.2dB gain of PSNR-l. The main reason can be concluded as that the PCD module enriches the feature representation ability, and aligns the features across multi-level features can handle more complex motions.

Dual Branches V.S. Single Branch We also conduct experiments to explore the effectiveness of the dual branches design. As shown in Table 3, the baseline model and its two variants are designed as a single branch, i.e., concatenating the LDR images and the gamma-corrected images directly. The results show that the dual branches design, which treats

the LDR images and their gamma-corrected images separately, is more effective, especially when compared with the baseline model AHDRNet.

5. Conclusion

We have presented ADNet, an attention-guided deformable convolutional network for multi-frame HDR imaging. A dual-branch pipeline is proposed where we handle the LDR images with a spatial attention module, and tackle misalignments with a PCD align module. Experimental results show that the proposed method can achieve state-of-the-art performance and reconstruct noise-free and ghost-free HDR images. Code used in this work will be publicly available upon publication.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No.61872067 and No.61720106004.

References

[1] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *Int. J. Comput. Vis.*, 92(1):1–31, 2011. 1

[2] Luca Bogoni. Extending dynamic range of monochrome and color images through fusion. In *Int. Conf. Pattern Recog.*, pages 7–12, 2000. 2

[3] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. *arXiv preprint arXiv:2009.07265*, 4, 2020. 3

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3

[5] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Conference on Computer Graphics & Interactive Techniques*, pages 369–378, 1997. 3

[6] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)*, 36(6):1–15, 2017. 1, 2

[7] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6):177, 2017. 2

[8] Orazio Gallo, Natasha Gelfandz, Wei-Chao Chen, Marius Tico, and Kari Pulli. Artifact-free high dynamic range imaging. pages 1–7, 2009. 2

[9] Thorsten Grosch. Fast and robust high dynamic range image generation with camera and object movement. *IEEE International Conference of Vision, Modeling and Visualization*, 2006. 2

[10] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1

[11] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. Hdr deghosting: How to deal with saturation? In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1163–1170, 2013. 2

[12] Katrien Jacobs, Celine Loscos, and Greg Ward. Automatic high-dynamic range image generation for dynamic scenes. *IEEE Computer Graphics and Applications*, 28(2):84–93, 2008. 2

[13] Takao Jinno and Masahiro Okuda. Motion blur free hdr image acquisition using multiple exposures. In *IEEE Int. Conf. Image Process.*, pages 1304–1307, 2008. 2

[14] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144, 2017. 1, 2, 3, 4

[15] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Trans. Graph.*, 22(3):319–325, 2003. 2

[16] Erum Arif Khan, Ahmet Oguz Akyuz, and Erik Reinhard. Ghost removal in high dynamic range images. In *IEEE Int. Conf. Image Process.*, pages 2005–2008, 2006. 2

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[18] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 596–611, 2018. 1

[19] Shuaicheng Liu, Ping Tan, Lu Yuan, Jian Sun, and Bing Zeng. Meshflow: Minimum latency online video stabilization. In *Eur. Conf. Comput. Vis.*, pages 800–815, 2016. 1

[20] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1651–1660, 2020. 1

[21] Kunming Luo, Chuan Wang, Shuaicheng Liu, Haoqiang Fan, Jue Wang, and Jian Sun. Upflow: Upsampling pyramid for unsupervised optical flow learning. *arXiv preprint arXiv:2012.00212*, 2020. 1

[22] Kede Ma, Zhengfang Duanmu, Hanwei Zhu, Yuming Fang, and Zhou Wang. Deep guided learning for fast multi-exposure image fusion. *IEEE Transactions on Image Processing*, 29:2808–2819, 2019. 1

[23] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pages 382–390. IEEE, 2007. 1

[24] Shree K Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 472–479, 2000. 1

[25] Tae-Hyun Oh, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Robust high dynamic range imaging by rank minimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(6):1219–1232, 2014. 2

[26] Fabrizio Pece and Jan Kautz. Bitmap movement detection: Hdr for dynamic scenes. In *Conference on Visual Media Production*, pages 1–8, 2010. 2

[27] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Aleš Leonardis, Radu Timofte, et al. NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 4

[28] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image hdr reconstruction using a cnn with masked features and perceptual loss. *arXiv preprint arXiv:2005.07335*, 2020. 1

[29] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.*, 31(6):203, 2012. 2, 3

[30] Jack Tumblin, Amit Agrawal, and Ramesh Raskar. Why i want a gradient camera. In *IEEE Conf. Comput. Vis. Pattern Recog.*, volume 1, pages 103–110, 2005. 1

[31] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 3, 6

[32] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Eur. Conf. Comput. Vis.*, pages 117–132, 2018. 1, 2, 3

[33] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1751–1760, 2019. 1, 2, 3, 5, 6

[34] Qingsen Yan, Dong Gong, Pingping Zhang, Qinfeng Shi, Jinqiu Sun, Ian Reid, and Yanning Zhang. Multi-scale dense networks for deep high dynamic range imaging. In *Proc. WACV*, pages 41–50, 2019. 2

[35] Qingsen Yan, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang. Deep hdr imaging via a non-local network. *IEEE Transactions on Image Processing*, 29:4308–4322, 2020. 2, 3

[36] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 3

[37] Wei Zhang and Wai-Kuen Cham. Gradient-directed multiexposure composition. *IEEE Trans. Image Process.*, 21(4):2318–2323, 2011. 2

[38] Henning Zimmer, Andrés Bruhn, and Joachim Weickert. Freehand hdr imaging of moving scenes with simultaneous resolution enhancement. In *Computer Graphics Forum*, volume 30, pages 405–414, 2011. 2