ATTENTION-MASK DENSE MERGER (ATTENDENSE) DEEP HDR FOR GHOST REMOVAL

Kareem Metwaly, Vishal Monga

School of Electrical Engineering and Computer Science
Pennsylvania State University, USA
Emails: kareem@psu.edu, vmonga@engr.psu.edu

ABSTRACT

High Dynamic Range (HDR) reconstruction is the process of producing an HDR image from a set of Standard Dynamic Range (SDR) images with different exposure times. This is a particularly challenging problem when relative camera or object motion exists between the available SDR images. Recently, deep learning methods, specifically those based on convolutional neural networks (CNNs) have been developed for HDR and shown to achieve unprecedented quality gains. Invariably an image alignment phase precedes the CNN mapping and merging. In practice, this alignment step greatly increases the computational burden of deep HDR methods often rendering them unsuitable for real-time composition. We propose a new deep HDR technique that does not need any explicit alignment of SDR images. Instead, a novel attention mask is developed that enables the network to focus on parts of the scene with considerable motion. Further, a dense merger is proposed that leads to an economical network. Evaluation over benchmark databases reveals that the proposed AttenDense network achieves high quality HDR results with significantly reduced computation time than state of the art. Further, the incorporation of domain knowledge (development of a custom attention mask) allows a more graceful decay in performance in the face of limited training.

1. INTRODUCTION

HDR images are desirable in several consumer imaging products but generally, expensive equipment is required to capture an HDR image. This has led to the idea of generating an HDR image out of a set of SDR images with different exposure times, thus relieving the need for such equipment [1, 2]. However, generating high quality HDR images continues to be a challenging problem. There has been a significant amount of research dedicated towards both HDR image [3, 4] and video HDR [5, 6, 7] composition.

Abstractly, the process starts with capturing N SDR images with different exposure times. This allows each image to capture different segments of the scene. In other words, for any segment of the scene, there exists at least one SDR image that is not saturated in those segments [8]. The key problem lies in merging those images to generate the desired HDR one.

Several problems occur naturally in this process. For instance, noise amplification can occur, where the noise in the image with short-time exposure is amplified in the process of normalizing various exposure times [9]. A particularly important problem is ghost artifacts, which happens when the input images are not geometrically consistent due to the existence of relative motion due to either camera or object movement.

Related work in HDR deghosting can be classified into two categories – recent learning based methods and traditional model-based methods . Extensive research has been devoted towards the removal of ghost artifacts via model-based methods which invariably involves a motion or correspondence estimation between SDR images prior to merging them using a weighted radiance and camera-response model [10]. For instance, Lee *et al.* and Oh *et al.* [3, 11] adopt rank minimization approach, by modeling the static scene as low-rank and motion as sparse. Hafner *et al.* [12] proposed a model to compute the HDR image with the optical flow simultaneously. A more comprehensive survey of traditional methods can be found in [13].

Learning based methods have emerged as a new wave in HDR image/video composition. CNNs are the most common deep networks employed [14, 15, 16]. Unprecedented performance gains have been reported with recently developed deep HDR methods [17]. Notable approaches in this direction include the work of Eilertsen *et al.* [15], which develops a neural network to construct an HDR image out of a single image. Kalantari *et al.* [18] also provide a new dataset with ground truth that can be used in training Convolutional Neural Networks (CNNs). Zhang and Lalonde [14] used a CNN to construct an HDR image from a panorama input. Deep networks achieve enhanced results due to two principal reasons. First, the availability of training data (many pairs of SDR images with a ground truth HDR image) helps learn more sophisticated mappings from SDR to HDR. Second, invariably an image alignment phase (typically based on optical flow [19, 20]) precedes the CNN mapping and merging. In practice, the presence of this alignment step greatly increases the computational burden of deep HDR methods often rendering them unsuitable for real-time composition [18]. To the best of our knowledge, most of CNN-based approaches perform optical flow alignment before merging except for [21]. Motivated

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

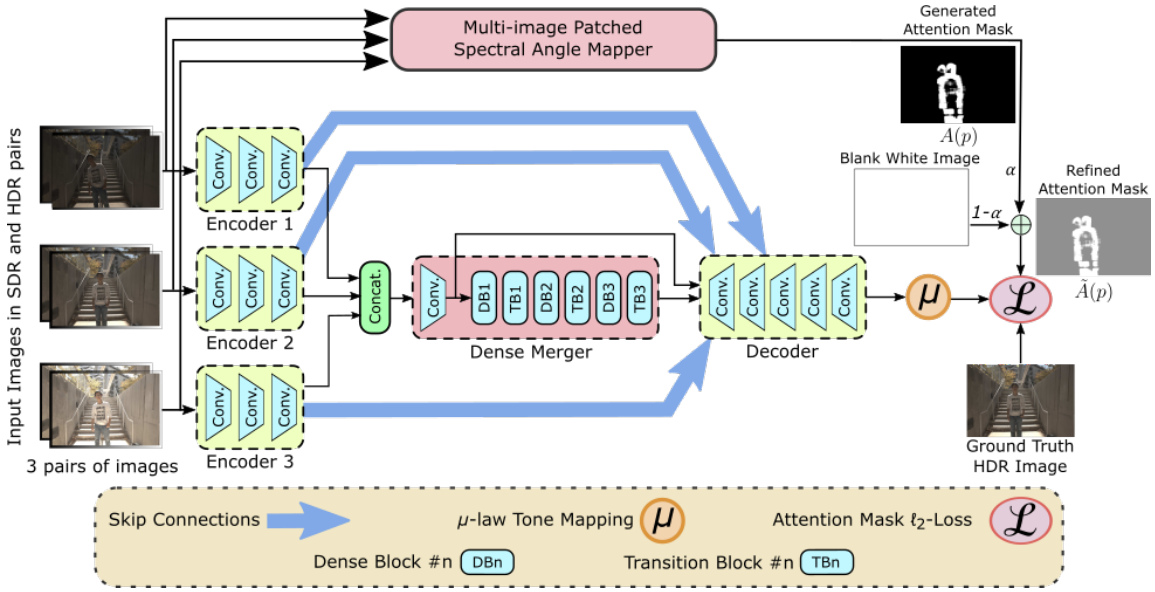

Fig. 1. AttenDense: Consists of three encoders, dense merger, decoder, tonemapper and attention mask synthesis.

by this, we develop a new AttenDense network that does not need an explicit alignment step. Instead, AttenDense focuses on computational efficiency and performance by the exploitation of scene characteristics towards formulation of a domain-enriched loss function in training the network. Specifically, our contributions are summarized as follows:

- Proposing a deep HDR solution that is significantly more efficient in terms of time and memory. This is done by mitigating the preprocessing phase for alignment and warping of input images.
- Utilizing domain knowledge to boost the performance and speed-up the learning process. This is performed by the development of an attention mask that enables the network to focus on more complicated regions. Subsequently, an attention mask weighted loss is employed in the training leading to superior results.
- Results on a benchmark database (that also includes training image pairs) are reported to confirm the computational cost vs. achievable HDR image quality benefits of AttenDense.

The paper is organized as follows. Section 2 discuss the architecture of the CNN, which we call AttenDense, and the process for generating attention masks. Experimental results are then presented in Section 3. Finally, we summarize the work and discuss potential future directions in Section 4.

2. ATTENDENSE CNN

2.1. Network Structure

The network architecture is depicted in Figure 1. There are three encoders for the three input exposures. However, the network is generic and the process can be extended for a larger number of input images. The objective of these encoders is to extract features out of different images that will be used later in merging. Each encoder takes six-channel input image. The first three channels correspond to the RGB channels of the SDR input image. The other three channels correspond to the HDR-transformed input image as follows,

$$H_i = S_i^{\gamma}/T_i \forall i \in 1,2,3 \tag{1}$$

where H_i is the transformed HDR input image, S_i is the input (3-channels RGB) SDR image, T_i is the exposure time and γ is the gamma correction factor which we set to 2.2 as in [18].

Each encoder consists of three blocks to increase the number of channels from 6 to 64, then 128 and finally 256. Each block consists of a convolution layer, a Leaky-ReLU activation function and batch normalization. All convolution layers have a stride of two and a filter kernel of size 5×5.

The output of the three encoders is concatenated and fed to the dense merger. The dense merger consists of two phases. The first phase performs a further encoding on the concatenated values. In the second phase, interlaced execution of dense blocks and transition blocks is carried out three times. Each dense block consists of two convolution blocks with ReLU activation functions. We use a kernel size of three and stride of one. The output of each dense block is the input that was fed to it plus the output of these two convolution blocks. Then a transition block recompresses the size of the output of a dense block to the original input size. Using a dense architecture enhances the flow of the gradient while allowing the network to get deeper [22].

There are skip-connections between respective blocks (blue arrows in Figure 1) in the encoders and the decoder to

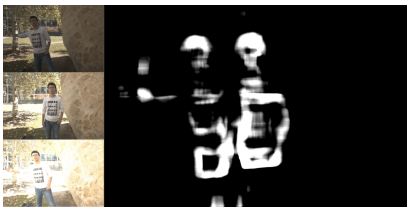

Fig. 2. Three input images and the generated attention mask. Note the attention mask has higher (close to white) values for regions with high motion and lower (close to black) values for static regions.

further enhance the flow of the gradient as well. In addition, it enables the decoder to retain some of the information that was lost during the encoding process. We perform transpose convolution operations in those decoding blocks with the same stride and kernel as in the encoder. The numbers of output channels in decoding blocks are $256, 128, 64, 64, 3$.

Tonemapper transforms the generated HDR image to an SDR one. Since we need a differentiable tonemapping, we used the μ-law tonemapper as in [18], which is defined as:

$$S = \frac{\log(1+\mu H)}{\log(1+\mu)} \tag{2}$$

where S is the SDR generated image, H is the original HDR image and $\mu = 5000$. Generally, this step emphasizes the importance of regions with low brightness; those will give higher gradient values.

We use l_2 distance to compute the loss, however we modify it by the attention mask to encourage the network design to focus on dynamic (high-motion) regions.

$$\mathcal{L}(\hat{y}, y) = \sum_{p \in pixels} \tilde{A}(p) \left(\hat{y}(p) - y(p)\right)^2 \tag{3}$$

where y is the ground truth tonemapped HDR image, $\hat{y}$ is the output of the network and $\tilde{A}(p)$ is the softened generated attention mask. By softened attention mask, we mean that for each pixel p,

$$\tilde{A}(p) = \alpha A(p) + (1-\alpha) \tag{4}$$

The softening in (4) enables a desirable balance between dynamic and static regions. We set $\alpha = 0.4$ by cross validation.

2.2. Attention Mask

The usage of an attention mask enables the network to perform well while training and maintaining the same level of complexity in testing as this block is removed after training. To generate the attention mask, we adopt the usage of Spectral Angle Mapper (SAM)[23]. The advantage of SAM is its invariance to changes in exposure time between images. In addition, it has a saturating behaviour for values that are close to 0 or 1 due to the usage of $\arccos$ function, thus reducing the effect of noise. SAM between two images I_1, I_2 is

$$\text{SAM}_{I_1,I_2}(p) = \arccos \frac{\langle I_1(p), I_2(p)\rangle}{\|I_1(p)\| \, \|I_2(p)\|}, \forall p \in pixels \tag{5}$$

where $< ., . >$ is the inner product operation and $\|I_1(p)\| = \sqrt{\langle I_1(p), I_1(p)\rangle}$. In our case of $N = 3$ images as inputs, we extended the definition to be

$$\text{SAM}_{\{I_n\}_{n=1}^3}(\bar{p}) = \max_{n \in \{1,2\}} \arccos \frac{\langle I_n(\bar{p}), I_{n+1}(\bar{p})\rangle}{\|I_n(\bar{p})\| \, \|I_{n+1}(\bar{p})\|} \tag{6}$$

where $\bar{p}$ corresponds to patches of the image. It is still feasible to extend the definition of SAM for a case of $N > 3$ by considering all different combinations of images and considering the maximum value over all of them. The reason for choosing the maximum value is to ensure that our network will focus on any region with any disruption.

An example for input images and their corresponding attention mask output is in Figure 2. We applied a limiting function as follows to clean the output.

$$h(x) = \begin{cases} 0 & x \le T_1 \\ \frac{x-T_1}{T_2-T_1} & T_1 < x \le T_2 \\ 1 & x > T_2 \end{cases} \tag{7}$$

T_1 and T_2 are data dependent and were chosen by cross-validation [24].

3. EXPERIMENTAL RESULTS

Experimental setup: We perform training and testing using Kalantari's dataset [18]. While there exist other datasets used for HDR deghosting, this is one of the very few that contains ground truth HDR images. This has indeed been the standard dataset for deep HDR for training and benchmarking [4, 18, 21]. This dataset contains 74 samples for training and 15 for testing. Each sample consists of 3 SDR input images, the exposure time of each SDR image and the corresponding HDR output. First, we report numerical evaluation (via PSNR and SSIM measures) of AttenDense compared with state of the art in model-based methods, as well as deep learning techniques for HDR. Next, we show a visual comparison between AttenDense, [18] and [21]. Finally, we compare AttenDense to a state of the art deep HDR method [21] without optical flow alignment in a limited training scenario.

3.1. Numerical Evaluation

We present an evaluation of the running time of different techniques in Table 1. The evaluation was performed on an Nvidia TITAN X GPU. AttenDense has the smallest running time. The reason is fundamentally because of removing the necessity for using optical flow alignment, which is typically an expensive pre-processing step. In addition, the attention mask is only used in the training phase, thus no additional overhead is incurred in the testing phase. Moreover, adopting dense layers in the merging enables reduction of layers significantly,

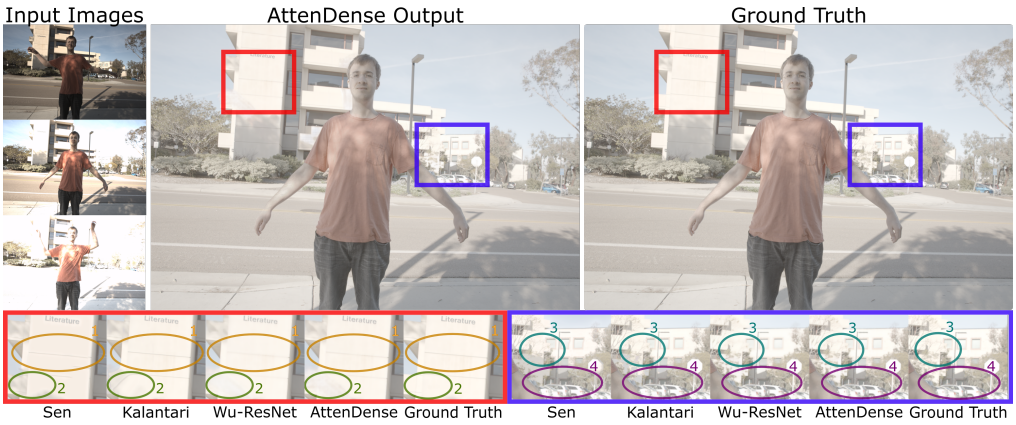

Fig. 3. Three input images and the corresponding generated output.

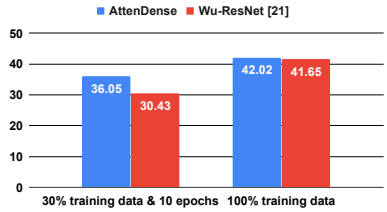

Fig. 4. Performance in Varying Training Scenarios.

Table 1. Average running time of AttenDense and state-of-the-art methods in model-based and Deep HDR

	Model-based Approaches		Deep HDR with Optical Flow	Deep HDR without Optical Flow	
	Sen [25]	Hu [26]	Kalantari[18]	Wu-ResNet [21]	AttenDense
Time (seconds)	58.6	30.8	16.2	3.3	**2.8**

which leads to an economical network in run time, number of network parameters (memory usage).

In Table 2, a performance evaluation of AttenDense compared to different methods is presented. AttenDense achieves state of the art performance and in particular beats competing deep HDR method(s) that also do not employ explicit alignment. AttenDense results are comparable to [18] with the latter producing mildly better PSNR/SSIM values. However, [18] network is more than **5.7x** slower than AttenDense, which is nearly real-time (see Table 1). Overall, AttenDense provides the most favorable cost-quality trade-off.

3.2. Visual Comparison

AttenDense consistently produces a more accurate image with greater amplitude detail than [21], [25]. Further, AttenDense can do better than [18] in regions with high motion where optical flow may be inaccurate and the following CNN exaggerates the motion error. In Figure 3, we show an example where, for instance, the horizontal lines (elliptical shape 1) are clearer in AttenDense and Kalantari's results. In elliptical shape 2, [18] suffers because of the hand motion and inaccurate optical flow alignment leading to a spurious artifact. Similarly, Sen *et al.* (which is a state of the art model based method) produces a shadow artifact in the region enclosed in elliptical shape 2. AttenDense is able to produce close to ground truth results. Similar arguments can be made for the highlighted regions in ellipses marked 3 and 4 respectively in Figure 3.

Table 2. Performance comparison between AttenDense and state-of-the-art methods in model-based and Deep HDR

	Model-based Approaches		Deep HDR with Optical Flow	Deep HDR without Optical Flow	
	Sen [25]	Hu [26]	Kalantari[18]	Wu-ResNet [21]	AttenDense
PSNR	40.80	35.79	42.70	41.65	42.02
SSIM	0.9808	0.9717	0.9877	0.9860	0.9870

3.3. Limited training scenario

To validate the performance of AttenDense, we compare the performance against [21] (state of the art in non-optical flow deep HDR) in case of using only 30% of the training set and employing just 10 epochs. Remarkably, with limited training, the benefits of AttenDense are even more pronounced. This is expected because the usage of attention mask serves as *additional domain specific information* enabling the AttenDense network to adapt better to paucity of training data.

4. CONCLUSION

We develop a new domain enriched deep HDR method called AttenDense to remove ghost artifacts in High Dynamic Range (HDR) image reconstruction. Compared to most state of the art deep HDR methods, our proposed AttenDense alleviates the need for explicit alignment as a pre-processing step, thereby leading to significant computational gains in practice. Instead, a novel attention mask is developed that enables the network to focus on parts of the scene with considerable motion. Further, a dense merger is proposed that leads to an economical network. Experiments over a recently designed and challenging benchmark dataset reveals that overall, AttenDense provides the most favorable cost-quality trade-off.

5. REFERENCES

[1] O. T. Tursun, A. O. Akyüz, A. Erdem, and E. Erdem, "The State of the Art in HDR Deghosting: A Survey and Evaluation," *Computer Graphics Forum*, vol. 34, no. 2, pp. 683–707, May 2015.

[2] S. Silk and J. Lang, "Fast high dynamic range image deghosting for arbitrary scene motion," in *Proc. of Graphics Interface*, 2012, pp. 85–92.

[3] C. Lee, Y. Li, and V. Monga, "Ghost-Free High Dynamic Range Imaging via Rank Minimization," *IEEE Signal Process. Letters*, vol. 21, no. 9, Sept. 2014.

[4] Q. Yan, D. Gong, P. Zhang, Q. Shi, J. Sun, I. Reid, and Y. Zhang, "Multi-Scale Dense Networks for Deep High Dynamic Range Imaging," in *2019 IEEE Winter Conf. on Applications of Comp. Vision*, Jan. 2019, pp. 41–50.

[5] S. Croci, T. O. Aydin, N. Stefanoski, M. Gross, and A. Smolic, "Real-time temporally coherent local HDR tone mapping," in *2016 IEEE Int. Conf. on Image Process.*, Phoenix, AZ, USA, Sept. 2016, pp. 889–893.

[6] Y. Li, C. Lee, and V. Monga, "A Maximum a Posteriori Estimation Framework for Robust High Dynamic Range Video Synthesis," *IEEE Trans. on Image Process.*, vol. 26, no. 3, pp. 1143–1157, Mar. 2017.

[7] N. K. Kalantari and R. Ramamoorthi, "Deep HDR Video from Sequences with Alternating Exposures," *Comp. Graph. Forum*, vol. 38, no. 2, pp. 193–205, 2019.

[8] K. Ma, H. Li, H. Yong, Z. Wang, D. Meng, and L. Zhang, "Robust Multi-Exposure Image Fusion: A Structural Patch Decomposition Approach," *IEEE Trans. on Image Process.*, vol. 26, no. 5, May 2017.

[9] M. Granados, K. I. Kim, J. Tompkin, and C. Theobalt, "Automatic Noise Modeling for Ghost-free HDR Reconstruction," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 201:1–201:10, Nov. 2013.

[10] Al Bovik, *Handbook of image and video processing*, Academic Press, 2000.

[11] T. Oh, J. Lee, Y. Tai, and I. S. Kweon, "Robust High Dynamic Range Imaging by Rank Minimization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1219–1232, June 2015.

[12] D. Hafner, O. Demetz, and J. Weickert, "Simultaneous HDR and Optic Flow Computation," in *2014 Int. Conf. on Pattern Recognition*, Aug. 2014, pp. 2065–2070.

[13] K. Karaduzovic-Hadziabdic, J. H. Telalovic, and R. K. Mantiuk, "Assessment of multi-exposure HDR image deghosting methods," *Computers & Graphics*, vol. 63, pp. 1–17, Apr. 2017.

[14] J. Zhang and J. Lalonde, "Learning High Dynamic Range from Outdoor Panoramas," in *2017 IEEE Int. Conf. on Comp. Vis.*, Venice, Oct. 2017, pp. 4529–4538.

[15] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep CNNs," *ACM Trans. on Graphics*, vol. 36, no. 6, pp. 1–15, Nov. 2017.

[16] F. Peng, M. Zhang, S. Lai, H. Tan, and S. Yan, "Deep HDR Reconstruction of Dynamic Scenes," in *IEEE Int. Conf. on Image, Vis. and Comput.*, 2018, pp. 347–351.

[17] S. Jia, Y. Zhang, D. Agrafiotis, and D. Bull, "Blind high dynamic range image quality assessment using deep learning," in *2017 IEEE Int. Conf. on Image Process.*, Beijing, Sept. 2017, pp. 765–769.

[18] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Trans. on Graphics*, vol. 36, no. 4, pp. 1–12, July 2017.

[19] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *The IEEE Int. Conf. on Computer Vision*, December 2015.

[20] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *The IEEE Conf. on Computer Vision and Pattern Recognition*, July 2017.

[21] S. Wu, J. Xu, Y. Tai, and C. Tang, "Deep High Dynamic Range Imaging with Large Foreground Motions," in *ECCV 2018*, vol. 11206, pp. 120–135.

[22] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, July 2017, pp. 2261–2269.

[23] M. M. Khan, "High Dynamic Range Image Deghosting Using Spectral Angle Mapper," *Computers*, vol. 8, no. 1, pp. 15, Mar. 2019.

[24] V. Monga, *Handbook of Convex Optimization Methods in Imaging Science*, Springer, 2018.

[25] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust patch-based hdr reconstruction of dynamic scenes," *ACM Trans. on Graphics*, vol. 31, no. 6, pp. 1, Nov. 2012.

[26] J. Hu, O. Gallo, K. Pulli, and X. Sun, "HDR Deghosting: How to Deal with Saturation?," in *2013 IEEE Conf. on Computer Vision and Pattern Recognition*, Portland, OR, USA, June 2013, pp. 1163–1170.