DEEP EXPOSURE FUSION WITH DEGHOSTING VIA HOMOGRAPHY ESTIMATION AND ATTENTION LEARNING

Sheng-Yeh Chen   Yung-Yu Chuang

National Taiwan University

ABSTRACT

Modern cameras have limited dynamic ranges and often produce images with saturated or dark regions using a single exposure. Although the problem could be addressed by taking multiple images with different exposures, exposure fusion methods need to deal with ghosting artifacts and detail loss caused by camera motion or moving objects. This paper proposes a deep network for exposure fusion. For reducing the potential ghosting problem, our network only takes two images, an underexposed image and an overexposed one. Our network integrates together homography estimation for compensating camera motion, attention mechanism for correcting remaining misalignment and moving pixels, and adversarial learning for alleviating other remaining artifacts. Experiments on real-world photos taken using handheld mobile phones show that the proposed method can generate high-quality images with faithful detail and vivid color rendition in both dark and bright areas.

Index Terms— Exposure fusion, deghosting, homography estimation, attention learning, adversarial learning.

1. INTRODUCTION

A camera is an imperfect device for measuring the radiance distribution of a scene because it cannot capture the full spectral content and dynamic range. High dynamic range (HDR) imaging and multi-exposure image fusion (MEF) are techniques to expand the low dynamic range (LDR) due to limitations of the camera sensor. Thanks to the development of imaging devices, we are able to capture a sequence of multi-exposure images in a short time to fulfil the dynamic range of a scene. With the sequence, HDR imaging methods try to recover the response curve and construct an HDR radiance map. MEF methods are applied to blend well-exposed regions from multiple images with different exposures to produce a single visually pleasing LDR image that appears to possess a higher dynamic range. Most MEF or HDR methods perform well with perfectly static sequences [1, 2, 3]. However, in practice, a certain amount of camera and object motions are inevitable, resulting in ghosting and blurry artifacts in the fused images.

Some techniques have been proposed for addressing the ghosting problem. While camera motion can be compensated

Fig. 1: An example result of the proposed method. We propose a learning-based approach to produce a well exposed LDR image given two differently exposed LDR images of a dynamic scene taken via a handheld mobile phone. We also compare our results with two previous methods by Ma et al. [2] and Prabhakar et al. [3].

using global alignment methods such as median threshold bitmap (MTB) [4], deghosting of moving objects is more difficult. Global image misalignment can also be compensated by homography [5]. Wu et al. [6] applied feature-based registration for estimating homography of multi-exposure images. However, due to parallax in saturated regions, the method cannot produce perfect alignment, and the final output may be blurry. Kalantari et al. [7] aligned exposure stacks using a traditional optical flow technique. Their method corrects the distortion owing to optical flow warping with a CNN, but still have artifacts for extreme dynamic scenes. Yan et al. [8] proposed an attention-guided network for ghost-free HDR imaging. However, their method requires a large amount of data with moving objects, and needs at least three images in the exposure stack for generating an HDR image.

This paper proposes a deep network for exposure fusion. For reducing potential ghosting, our network is retrained to take only two images, an underexposed image and an over-

exposed image. The network consists of three main components: a homography estimation network for compensating camera motion, a fusion network with attention learning for reducing misalignment and moving pixels, and a discriminator network for adversarial learning, which alleviates remaining artifacts. Our contribution are as follows.

- We propose the first deep network for estimating a homography between two differently exposed images and integrate it into an exposure fusion network. Previous methods are either designed for images with similar exposures or based on conventional methods.
- We present a model that produces a ghost-free image from two images with large exposure difference via attention learning.
- To the best of our knowledge, we are the first to apply adversarial learning for exposure fusion.

2. METHOD

Given an underexposed image I_1 and an overexposed image I_2 of a scene, our goal is to generate an image with better color and structure rendition in both dark and bright regions. Since nowadays photos are often taken with handheld cameras, it is necessary to handle both camera and scene motion. Our method has two stages: alignment and fusion. During alignment, the overexposed image I_2 is registered to the underexposed image I_1. The aligned images are then blended into a visually pleasing image at the fusion stage.

Our method addresses both alignment and fusion problems using a convolutional neural network (CNN). Fig. 2 gives an overview of our method. Given the underexposed image I_1 and the overexposed image I_2 as the input, first, the homography network estimates the homography that warps I_2 to align with I_1. Next, the warped overexposed image I_2^w and the underexposed image I_1 are fed to the generator to produce an image with better visual quality. Finally, the discriminator takes the underexposed image I_1, the warped overexposed image I_2^w, and the reference image I_r or the generated result I_f as input and predicts if the third image is the reference image or the generated image. This way, the generator is trained to produce images indistinguishable with the reference images, further improving the visual quality.

For training the network, we use the SICE dataset collected by Cai et al. [9]. The SICE dataset contains over 500 exposure stacks of static scenes with high-quality reference images. The reference images were selected after pairwise comparisons of results generated by 13 state-of-the-art multi-exposure fusion or stack-based HDR methods at that time by 13 amateur photographers and 5 volunteers.

2.1. Homography network

4-point homography parameterization. A homography is often represented as a 3×3 matrix H_{matrix} which maps a
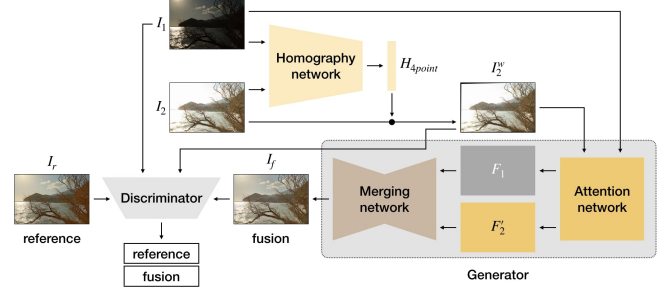

Fig. 2: The overview of the proposed method.

pixel $[u, v]$ of the source image to the pixel $[u', v']$ of the destination image by matrix multiplication so that they are aligned. However, as pointed out by DeTone et al. [10], it is difficult to balance the rotational and translation terms by using such a representation in an optimization problem. They suggest to use an alternative 4-point parameterization H_{4point} by using the offsets of the four corner locations, $\Delta u_i = u'_i - u_i, \Delta v_i = v'_i - v_i$ for $i = 1, ..., 4$. We adopt this representation because it leads to more stable optimization.

Training example generation. Since images in the SICE dataset [9] were captured for static scenes with tripods, the images of a scene are well aligned although with different exposures. For preparing training data for the homography network that predicts H_{4point}, we induce a random projective transformation to a pair of aligned images at a time. This way, we have an unlimited number of training examples.

Homography network N_h. The objective of the alignment stage is to warp the overexposed image I_2 to the underexposed image I_1. For this purpose, we construct a homography network which is composed of 3×3 convolutional blocks with instance normalization [11] and leaky ReLUs. The images are resize to 256×256 as the predicted range of the offsets must be fixed, and the median threshold bitmaps (MTB) of the images are taken as augmented input. The network has 12 convolutional layers with a max pooling layer (2×2, stride 2) after every two convolutional layers. The numbers of filters are 64 for the first four layers, 128 for the next four layers, and 256 for the last four layers. At the end, there are two fully connected layers with 1024 and 8 units. Fig. 3(a) depicts the architecture of the homography network. The network produces eight real numbers for H_{4point} and is trained using an L_2 loss. We then applied the homography for warping I_2 into its aligned version, I_2^w. Eventually, we have $[H_{4point}, I_2^w] = N_h(I_1, I_2)$.

2.2. Generator

Given the registered images I_1 and I_2^w, the generator attempts to fuse them into a visually pleasing image. As shown in Fig. 3(b), our generator consists of two modules: the attention network and the merging network.
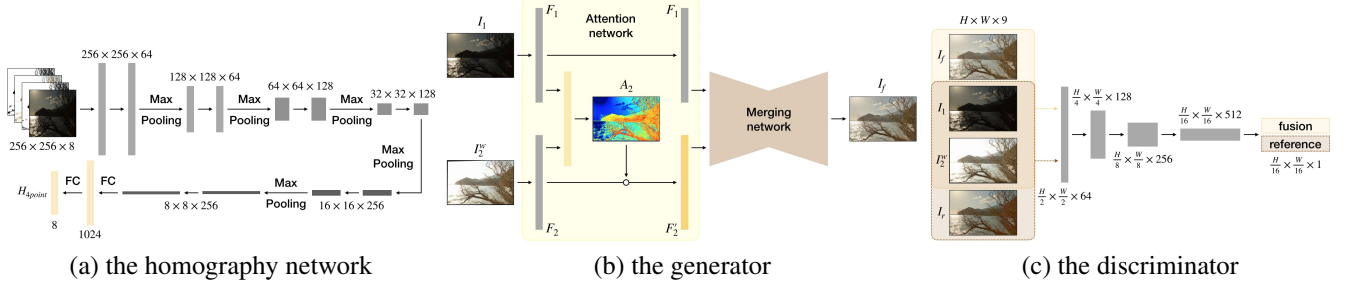
(a) the homography network   (b) the generator   (c) the discriminator

Fig. 3: The architecture of the sub-networks, the homography network, the generator and the discriminator.

Attention network N_a. Inspired by Yan' et al. [8], we use an attention mechanism on the warped overexposed image I_2^w to highlight the well-aligned area for generating ghost-free fusion results. Given I_1 and I_2^w, the attention network first uses a shared encoding layer to extract feature maps F_1 and F_2 with 64 channels. F_1 and F_2 are then sent to the convolutional attention module $a_2(\cdot)$ for obtaining the attention map A_2 for I_2^w, i.e., $A_2 = a_2(F_1, F_2)$. The attention module has two convolutional layers with 64 3×3 filters, respectively followed by a leaky ReLU activation and a sigmoid function. The predicted attention map is used to attend the features of the warped overexposed image by $F_2' = A_2 \circ F_2$, where $\circ$ denotes the element-wise multiplication and F_2' is the feature map after attention. As a result, the attention network N_a takes I_1 and I_2^w from the homography network and generates F_1 and F_2': $[F_1, F_2'] = N_a(I_1, I_2^w)$. The attention maps can suppress the misaligned and saturated regions in the overexposed image, avoiding the unfavorable features getting into the merging process and alleviating ghosting accordingly.

Merging network N_m. We model the merging network after U-Net [12] with seven downsampling convolutional layers followed by seven upsampling deconvolutional layers. Each layer is composed of 64 4×4 filters. By taking inputs from the attention network, F_1 and F_2', the merging network produces an image I_f, exhibiting abundant color and structure information in both dark and bright regions by $[I_f] = N_m(F_1, F_2')$.

The optimization target of the merging network is to minimize the difference between the fused image I_f and the reference image I_r. For measuring differences, we use the perceptual loss that has been successfully applied to tasks such as image synthesis [13], super-resolution [14], and reflection removal [15]. We obtain the perceptual loss by feeding I_f and I_r through a pretrained VGG-19 network Φ. We compute the L_1 loss between $\Phi(I_f)$ and $\Phi(I_r)$ for the layers 'conv1_2', 'conv2_2', 'conv3_2', 'conv4_2', 'conv5_2', and the 'input' layer in VGG-19:

$$L_{feat} = \sum_{I_1, I_2} \sum_l \lambda_l \parallel \Phi_l(I_r) - \Phi_l(I_f) \parallel_1, \qquad (1)$$

where Φ_l represents the l-th layer in the VGG-19 network. We assign the weighting terms λ_l as $\frac{1}{2.6}$, $\frac{1}{4.8}$, $\frac{1}{3.7}$, $\frac{1}{5.6}$, $\frac{1}{0.15}$ for the convolutional layers and 1 for the input layer [15].

2.3. Discriminator

The fused image produced by the generator could still suffer from undesirable color degradation and halo effects. For further improving the visual quality, we adopt the idea of adversarial learning by modeling after the conditional Patch-GAN [16]. We construct our discriminator N_D with five convolutional layers with 4×4 convolutional blocks. The numbers of filters are 64, 128, 256, 512, and 1. We use instance normalization for the middle three layers. Except the final convolutional layer, the stride number is 2 and the activation function is leaky ReLU. Fig. 3(c) shows the architecture. The discriminator attempts to discriminate between patches I_r from the reference images and patches I_f generated by N_m conditioned on I_1 and I_2. With the discriminator, the generator is trained to produce results more similar to the reference.

The adversarial loss is defined as LSGAN [17]. The optimization goal of the discriminator N_D is:

$$\arg\min_{N_D} \sum_{I_1, I_2} \frac{1}{2}[N_D(I_1, I_2, I_f)^2 + (N_D(I_1, I_2, I_r) - 1)^2], \qquad (2)$$

where $N_D(I_1, I_2, x)$ outputs the probability that x is a high quality fused image given the input images I_1, I_2. In addition, our adversarial loss L_{adv} is:

$$L_{adv} = \sum_{I_1, I_2} (N_D(I_1, I_2, I_f) - 1)^2. \qquad (3)$$

The goal of the generator is to minimize L_{feat} and L_{adv}:

$$\arg\min_{N_a, N_m} \sum_{I_1, I_2} w_1 L_{feat} + w_2 L_{adv}, \qquad (4)$$

where w_1=1 and w_2=0.01 are parameters balancing losses.

3. EXPERIMENTS

We first train our homography network N_h for 500 epochs, and then train the remaining networks (N_D, N_a, N_m) for 200 epochs. The whole training process takes about 27 hours on a single P100 GPU. We use a batch size of 4 patches and each epoch has 471 iterations. We optimize our networks using Adam optimizer [18] with a learning rate of 0.0001.

Fig. 4: Comparison with Ma et al. [2] and DeepFuse [3].

Fig. 5: Comparison of homography estimation with MTB.

Fig. 6: Ablation study. (a) w/o homography network (b) w/o attention mechanism.

For training, we use the training set of SICE dataset by Cai et al. [9]. All images are resized to 1200×800. Since our approach requires two differently exposed images as input, we sort all exposure stacks in the dataset by luminance, and take the i-th and the $N - i + 1$-th image in an exposure stack which contains N images as a training example, where $i = 1, ..., \frac{N}{2}$. We generate image patches of the size 256×256 in run time related by a 4-point homography parameterization randomly sampled by `truncated_normal()` in TensorFlow. We perturbed each corner of the 256×256 patch by a maximum of one eighth of the image width. We avoid larger random disturbance for avoiding extreme transformation which might affect the quality of fusion. We use the testing set of SICE dataset as the validation set to monitor overfitting. As for the testing set, we test our method on photos taken using a handheld mobile phone with the exposure bracketing mode. To test the robustness of our approach, the exposure values are selected automatically by the app.

Homography estimation. Fig. 5 compares our homography network with MTB [4]. On the bottom right, we show the XOR difference, indicating differences after alignment. Our result has less misaligned pixels than MTB as shown in the XOR difference. Our success rate is about 86.63% for 60 pairs of photos taken by handheld phone cameras.

Exposure fusion. Fig. 4 compares our method with the patch-based approach by Ma et al. [2] and the DeepFuse approach by Prabhakar et al. [3]. Note that we do not have ground-truth images here as the images were captured by handheld

cameras. The method of Ma et al.requires more images for generating reasonable results. With two images, their results exhibit color distortion and halo artifacts. DeepFuse requires static scenes and their results are blurry in these examples.

Ablation study on the architecture. For investigating the contribution of individual components, we compare variants of the proposed network, *w/o N_h* and *w/o N_a*. As shown in Fig. 6, the variant *w/o N_h* fails to align well when there is camera motion or moving objects, leading to blurry results. Although the *w/o N_a* variant can produce sharper results, it still has ghosting artifacts in saturated regions. The full model can alleviate ghosting artifacts and provide sharper details.

4. CONCLUSIONS

This paper presents a deep learning method for exposure fusion. For reducing ghosting artifacts and rendering sharp details, our model integrates homography estimation for compensating camera motion, attention mechanism for reducing influence of moving objects and adversarial learning for further improving visual quality. With these modules together, our method can generate images with vivid color and sharp details in both dark and bright areas from a pair of underexposed and overexposed images.

5. REFERENCES

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

[2] Kede Ma, Hui Li, Hongwei Yong, Zhou Wang, Deyu Meng, and Lei Zhang, "Robust multi-exposure image fusion: a structural patch decomposition approach," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2519–2532, 2017.

[3] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs.," in *ICCV*, 2017, pp. 4724–4732.

[4] Greg Ward, "Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures," *Journal of graphics tools*, vol. 8, no. 2, pp. 17–30, 2003.

[5] Zygmunt L Szpak, Wojciech Chojnacki, and Anton van den Hengel, "Robust multiple homography estimation: An ill-solved problem," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2132–2141.

[6] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang, "Deep high dynamic range imaging with large foreground motions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 117–132.

[7] Nima Khademi Kalantari and Ravi Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes.," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 144–1, 2017.

[8] 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," *arXiv preprint arXiv:1904.10293*, 2019.

[9] Jianrui Cai, Shuhang Gu, and Lei Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.

[10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Deep image homography estimation," *arXiv preprint arXiv:1606.03798*, 2016.

[11] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[13] Qifeng Chen and Vladlen Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1511–1520.

[14] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[15] Xuaner Zhang, Ren Ng, and Qifeng Chen, "Single image reflection separation with perceptual losses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4786–4794.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[17] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2794–2802.

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