HDR Reconstruction By Novel Framework

*Note: Group 33

1st Yen-Ting Huang
Group 33
ECE 228
University of California, San Diego
yeh001@ucsd.edu

2nd Ziyu Lin
Group 33
ECE 228
University of California, San Diego
zil332@ucsd.edu

3rd Huilin Song
Group 33
ECE 228
University of California, San Diego
hus018@ucsd.edu

***Abstract*—Due to different lighting and exposure, a photo may result in unexpected or uncleared representation, which can cause some eliminations of important aspects and objects. High-dynamic range(HDR) imaging is a critical research topic that aims to fuse images with different exposure to improve image quality in terms of illumination, and it is also widely employed to applications across multiple domains, including smartphone devices, unmanned aerial vehicles, autonomous driving system, etc. In this research, a novel knowledge distillation mechanism is applied to boost the performance of a HDR model. By incorporating the knowledge distillation mechanism to guide the training process of the HDR model, we can expect that a superior performance can be achieved. To our best knowledge, this is the first work that applying knowledge distillation to the task of image fusion.**

I. Introduction

Fig. 1. Examples of HDR imaging [1]

As shown in Fig. 1, images shot in different lighting conditions may lose details in those under- and over-exposure local patches. The degraded images may not only influence viewers' experience but also make poor predictions for computer vision applications, including object detection, semantic segmentation, visual odometry, scene understanding, etc. Several studies [2, 3] regarding image fusion for boosting poorly illuminated images have been developed. Enhancing luminance and contrast of images with end-to-end convolutional neural networks can give the most detailed information to the viewer. We noticed that knowledge distillation is a useful technique that has been applied to the tasks of image classification, object detection, image dehazing. We attempt to apply knowledge distillation to teacher and student training models and allow the student model to learn the best lighting and exposure of a photo, and then construct the clear photo that can provide well-illuminated images to the viewer. The input to our model is 2 photos taken from the same angle and with different exposure. We then use knowledge distillation neural network to output an image. The output image contains the most complete details of the scene, and it is adjust to the exposure level that is most suitable for human to visualize. This problem is important to be solved because it can offer people the most clear view of any surroundings.

II. Related work

A. Exposure Correction

Traditionally, people perform exposure correction by using graphics editor like Photoshop. However, there are researches about using machine learning to handle perform this task. [10] proposes an HDR model based on GAN to handle the object motions in the images. The adversarial learning framework is applied to generate high-quality image patches for those missing contents. [4] design a unified unsupervised image fusion network for multiple image fusion tasks, which includes multi-exposure fusion. It can adaptively fuse images according to the importance of the inputs. [3] also presents an unsupervised image fusion network based on a densely connected network. By applying elastic weight consolidation (EWC) to retain features from previous tasks. Although these approaches are effective and perform well on their corresponding data, it appears that they ignore the exposure degradation from all of high-exposed, low-exposed, and LDR images. [1] designs a coarse-to-fine deep neural network (DNN) model by laplacian pyramid decomposition that addresses each subprobem separately and achieves state-of-the-art performance. However, their qualitative results show that image details may lose. In our framework, we can handle the images captured from any two exposure degrees and generate more detaied results, with low-level features concatenation, than [1].

B. Knowledge Distillation

In recent years, researchers[5] applied knowledge distillation techniques to train the dehazing model. The technique would have a reference output from the teacher model, and the student model needs to dehaze the degraded images by the reference output, where the reference feature outputs provides guidance to the students during training. While inferencing, the teacher model is not used to generate clear images. It applies Feature-Based Knowledge to explicitly associate the intermediate-level representation of the teacher and student

model. Inspired by this work, we curious that whether this technique can also be introduced to the tasks of image fusion like HDR. Thus, we apply this framework as the reference model to develop our approach.

C. Attention Mechanisms

SENet(Squeeze-and-excitation networks)[12] addresses the classification problem by an effective Squeeze-and-excitation (SE) module. To be more specific, the SE module contains an extra branch for re-calibrating channels of output feature maps, which is a two-layer perceptron to catch global information. By introducing this module to different tasks, it can improve the performance for segmentation[13]. Strip Pooling Network(SPNet)[15] designs a novel and lightweight attention module to capture the long-term dependencies based on different pooling operations. The ablation study exhibits that strip pooling modules can build long-term contextual information and provide better segmentation results on popular scene segmentation datasets (Cityscapes and ADE20K). Hence, considering the efficiency and memory constraint, we prefer to apply the squeeze-and-excitation module and strip pooling module in this research.

III. Dataset

In the experiments, we use the dataset from [1][1]. This dataset was rendered from linear raw-RGB images taken from the MIT-Adobe FiveK dataset[17]. Each image was rendered with -1.5, -1, +0, +1, and +1.5 relative exposure values (EVs) by an emulation of the camera ISP processes. The dataset can be split into three sets: the training set of 17,675 images, the validation set of 750 images, and the testing set of 5,905 images. We use standard deviation normalization to process our input. The resolution of the images is 512×512. In our testing results, we will take all the combinations of the paired data in testing images to evaluate the performance. Thus, the number of the paired image with low exposure, high exposure, all exposure are 1181, 3543, and 11810, respectively.

IV. Methods

A. Teacher Model

An autoencoder framework is applied to distill the features of properly exposed images. As shown in Fig 2. The network takes in a ground truth image and reconstructs the input. There are three modules in the Teacher model, which are downsampling module, Resblock module, and upsampling module, respectively. For the downsampling module, we use convolution layers with stride two to halve the input. Two sets of Conv2d-ReLu are formed to downsample four times of the input. Then, six Resblocks proposed by [16] are introduced to build the latent feature representations for the knowledge distillation. We can consider that these encoded features exhibit different latent representations from the low level to the high level. Finally, the upsampling module, which is symmetric to the downsample module, would recover the input shape by two sets of ConvTranspose2d-ReLu.

[1]Dataset Link: Click Here

B. Student Model - Image Fusion Model

Inspired by [5], we construct a new fusion framework to synthesize multi-exposed images(I) and produce a well-illumination image(J). The overview of the framework is illustrated in Fig.2. Two images are concatenated and passed to the base layer. The learning rate for model training is 0.001, and the optimizer is Adam. The base layer is a two-layer convolution layer for extracting the low-level features. Then, two convolutions with stride 2 would downsample the feature maps. We adopt six Resblock to extract high-level features. After that, two mixed pooling modules, which are a part of the Strip pooling module, would capture long-range correlations. For the decoder, we use two upsampling to recover the resolution. To better preserve the image's details, we concatenate the low-level features from previous outputs. Then, a channel-wise attention module, which is the squeeze-and-excitation module in SENet, is used to reweight channel-wise features. Finally, we refine the output by a reconstruction layer. This layer has two convolution layers followed by a Tanh layer.

There are mainly two components that significantly improve the performance of our model. 1) Architecture: we concatenate the low-level features to the decoder to enhance the detail features. The model discussed in [5] can not generate high-quality images because the reconstruction of low features, such as texture and edges, would be lost. 2) Attention: We introduce the channel and spatial attention to our fusion model (Student model), so we can capture the correlation between channels and locations. The channel attention is applied to fuse and boost the important features from low-level features and high-level features. Mixed pooling modules from Strip Pooling are responsible for discovering semantic information by capturing the long-range contextual information. Thus, our model makes some improvement on those aspects.

We apply knowledge distillation techniques to train the model. Different from previous works that take a single input image to enhance its visual quality, knowledge distillation can be treated as guidance for the fusion process of two images. To be more specific, the strength of knowledge distillation is to use trained features from the teacher model to refine the student model for preventing gradient descent from converging to a local minimum. Our teacher model is an autoencoder model. It would provide intermediate feature representations of properly exposed images. We use the second, fourth, and sixth outputs of Resblock to guide our fusion model.

C. Loss Function

1) Loss function for the teacher model: The loss function of our teacher model is following:

$$L_{teacher} = ||J - T(J)||_1 \tag{1}$$

where T denotes the teacher model, J is the ground truth. Similar to existing works for autoencoder, we use the L1 loss to compute the difference between the predictions and the ground truth. The teacher model is proposed to generate the

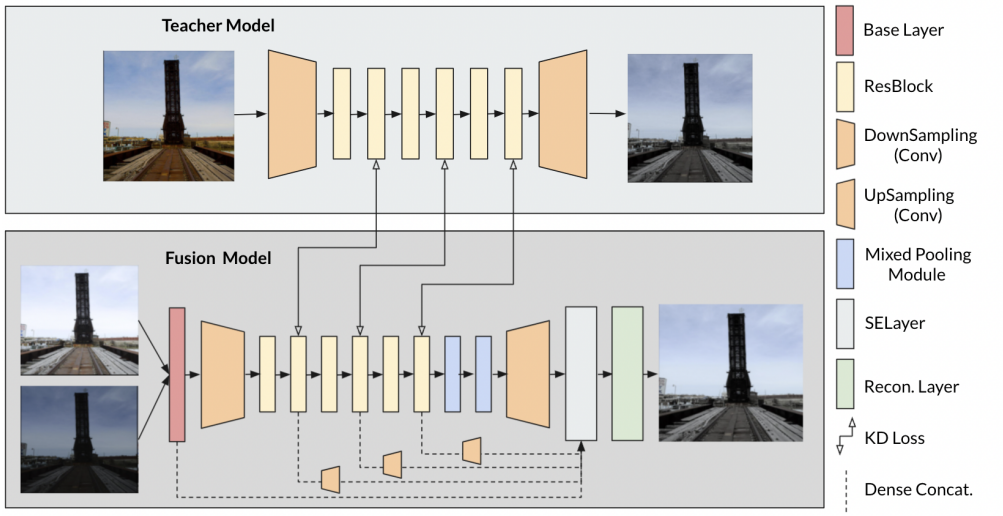

Fig. 2. An overview of the knowledge distilling dehazing network

guidance of feature maps that can teach the fusion model the features of properly exposed images.

2) Loss function for the fusion model: The loss function of our fusion model that consider the L1 loss, knowledge distillation, and perceptual loss is adopted as:

$$L_{fusion} = L_{rec} + \lambda_1 \cdot L_{kd} + \lambda_2 \cdot L_{perc} \tag{2}$$

where L_{rec} denotes the reconstruction loss, L_{kd} is the knowledge distillation loss, and L_{perc} represents the perceptual loss. λ_1 and λ_2 are hyperparameters to balance the influence of the individual losses. The individual losses are defined next.

- Reconstruction Loss

$$L_{rec} = ||J - S(I)||_1 \tag{3}$$

where S denotes the student model, J is the ground truth, and I is the input images. The reconstruction loss, which uses L1 loss, computes the absolute value of the difference between the prediction results of the student model and the ground truth. The reason we choose L1 loss function rather than L2 loss is that if the noise/outliers are present in images, they would lead to much more error for L2 loss. The images may contain a small amount of noise if one of the images is degraded. Hence, L1 loss is preferable for us to reconstruct the well-exposed images. We use L1 loss to train both the teacher model and the student model.

- Knowledge Distillation Loss

$$L_{kd} = \sum_{(m,n)\in C} ||S^m(I) - T^n(J)||_1 \tag{4}$$

where T denotes the teacher model. Knowledge distillation loss would teach the fusion model to mimic the feature distribution of the clean images. Therefore, it will measure the error between the feature maps of the teacher model and our fusion model. In our implementation, m and n are the output of the feature maps at the second, fourth, and sixth Resblock.

- Perceptual Loss

$$L_{perc} = \sum_{i=1}^{n} ||\phi^i(S(I)) - \phi^i(J)||_1 \tag{5}$$

where ϕ denotes the output from the i-th convolution layer (after activation) that i is at the layer of 2, 7, 12, 21, and 30 of VGG19. Perceptual loss uses the pre-trained 19 layer VGG network to measure the difference between our results and ground truth. It is often used in restoration tasks to ensure the consistency for features of both high-level features and low-level features.

V. Experiments/Results/ Discussion

A. Quantitative analysis

For comparison, the "AE" denotes the autoencoder/Teacher Model. The "KD Baseline" model represents the model from [5]. We perform image channel transformation, then train the model for the Y channel. For the "Final" model, we add all components introduced in Sec. 3, and experimental results on the Exposure-Errors Dataset show the proposed method(Final) can achieve a significant 1.8% increase in PSNR and 11% increase in SSIM compared with the original paper.

Using knowledge distillation on image fusion, we observe the model of "KD Baseline" achieves better PSNR and SSIM than previous SOTA results. For PSNR, we can improve approximately 1.3 higher. Our KD Baseline model can accomplish 9.3% higher SSIM than previous work, which is much evident compared to other metrics. The result shows that two image fusion with knowledge capture more structure and texture features to get well-illumination results.

By incorporating both low-level features to the decoder and channel-wise attention for high-level and low-level features,

Model	Data	PSNR (AE)	PSNR(Pred)	SSIM(AE)	SSIM(Pred)
SOTA[1] (CVPR 2020)	Low exposure	-	20.542	-	77.0
SOTA[1] (CVPR 2020)	High exposure	-	19.980	-	76.8
SOTA[1] (CVPR 2020)	All exposure	-	20.205	-	76.9
Ours(KD Baseline)	Low exposure	25.470	20.899	90.1	85.3
Ours(KD Baseline)	High exposure	27.122	21.713	90.8	86.5
Ours(KD Baseline)	All exposure	26.452	21.494	90.9	86.2
Ours(DenseFusion)	Low exposure	25.243	21.133	89.4	86.8
Ours(DenseFusion)	High exposure	26.747	22.244	89.3	87.9
Ours(DenseFusion)	All exposure	26.137	21.947	89.3	87.8
Ours(Final)	Low exposure	24.949	21.423	87.3	87.0
Ours(Final)	High exposure	26.353	22.082	87.2	87.9
Ours(Final)	All exposure	25.784	22.057	87.2	87.9

TABLE I
COMPARISON OF THE PERFORMANCE OF DIFFERENT EXPERIMENTAL SETTINGS. ∗ INDICATES THE BEST RESULT.

the results of the DenseFusion model can accomplish close scores compared with the Final model. We use the attention mechanism from Strip Pooling to associate the long-range features with high-level features. It can only give 0.1% improvements in PSNR, which may cause the PSNR decrease in high exposure images. For low-exposed images, the spatial attention mechanism can provide more contextual information and results in higher scores.

B. Qualitative Analysis

Fig. 3-5 includes five columns of images. The left two are the input images. The third and fourth images are the restored picture of the teacher model and fusion model. The last image is the ground truth. Results are shown for three cases: low-exposure images, high-exposure images, both low-exposure and high-exposure images, respectively. We solved two special conditions: extremely high dynamic range and extremely low dynamic range of low-exposed and high-exposed images. We also solve the processing of extremely dark images as well as a mixture of bright and dark areas.

We have two limitations to our proposed approach: the presence of tone mapping by the ground truth and extremely dark/bright for both input images. Experts will adjust the saturation and the chrominance for ground truth, but we cannot produce the color outside the chrominance range of the two images. Also, if the images are too dark or bright, the small signal becomes difficult to correct. For these two issues, we may use GAN-based techniques to overcome them in the future.

Fig. 3. Qualitative results of low-exposure images

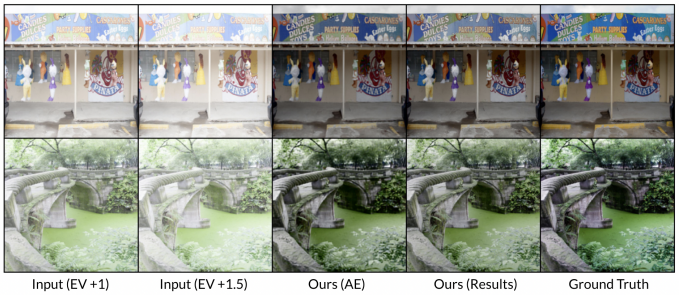

Fig. 4. Qualitative results of high-exposure images

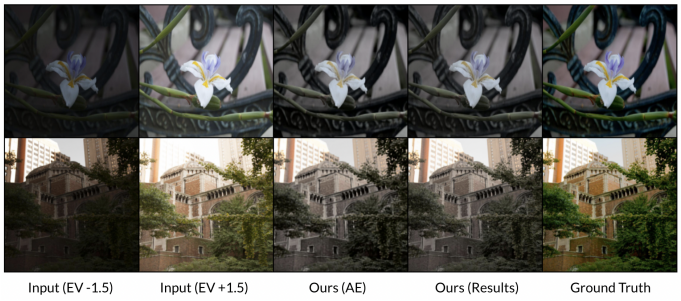

Fig. 5. Qualitative results of low-exposure and high-exposure images

VI. CONCLUSION

In this research, we propose an image fusion approach based on the knowledge distillation and attention mechanism for recovering image details from under-exposed or over-exposed areas in the input images. We also concatenate low-level features to avoid the loss of image details. Experimental results on the Exposure-Errors Dataset exhibit that the proposed method can achieve a significant 1.8% increase in PSNR and 11% increase in SSIM compared with the original paper. In the future, we plan to improve our model by GAN-based method to overcome the limitations of tone mapping and the loss of details.

CONTRIBUTIONS

- **Yen-ting Huang**: Software Developer, proposed the project topic, worked on developing the Teacher and student network.

- **Ziyu Lin**: Software Developer, working on training the data. Also in charge of documentation.
- **Huilin Song**: Software Developer, working on testing and tuning the hyperperameter. In charge of documentation.

This project shares dataset and infrastruction with project from ECE285.

REFERENCES

[1] Afifi, Mahmoud, et al. "Learning Multi-Scale Photo Exposure Correction." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[2] Bychkovsky, Vladimir, et al. "Learning photographic global tonal adjustment with a database of input/output image pairs." CVPR 2011. IEEE, 2011.

[3] Xu, Han, et al. "Fusiondn: A unified densely connected network for image fusion." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 07. 2020.

[4] Xu, Han, et al. "U2Fusion: A unified unsupervised image fusion network." IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[5] Hong, Ming, et al. "Distilling Image Dehazing With Heterogeneous Task Imitation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[6] R. A. Hummel, "Image enhancement by histogram transformation," Computer Graphics and Image Processing, vol.6, no.2, pp. 184-195, 1977.

[7] K. Zuiderveld, "Contrast Limited Adaptive Histograph Equalization," Graphic Gems IV., San Diego: Academic Press Professional, pp. 474-485, 1994.

[8] Li, Chongyi, et al. "LightenNet: A convolutional neural network for weakly illuminated image enhancement." Pattern Recognition Letters 104 (2018): 15-22.

[9] Wei, Chen, et al. "Deep retinex decomposition for low-light enhancement." arXiv preprint arXiv:1808.04560 (2018).

[10] Niu, Yuzhen, et al. "HDR-GAN: HDR image reconstruction from multi-exposed ldr images with large motions." IEEE Transactions on Image Processing 30 (2021): 3885-3896.

[11] Gou, Jianping, et al. "Knowledge distillation: A survey." International Journal of Computer Vision 129.6 (2021): 1789-1819.

[12] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[13] Zhong, Zilong, et al. "Squeeze-and-attention networks for semantic segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[14] Fu, Jun, et al. "Dual attention network for scene segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[15] Hou, Qibin, et al. "Strip pooling: Rethinking spatial pooling for scene parsing." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[16] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[17] Bychkovsky, Vladimir, et al. "Learning photographic global tonal adjustment with a database of input/output image pairs." CVPR 2011. IEEE, 2011.

Replies to critical reviews

Critical review from team 21:

1. How about using a variational autoencoder instead of an autoencoder?
   Our response: From what we learn, the purpose of autoencoder is to compress the input data, and the purpose of the variational autoencoder is to generate new data that's related to the original source data. Since our purpose is to compress the input data for our knowledge distillation model, we chose to use Autoencoder instead of variational autoencoder.

2. It is not clear why you used attention on image restoration.
   Our response: The purpose of our attention model is to enhance the important features, and weaken the effect of the trivial features. Channel attention is used to modelling interdependencies between channels. It is composed of SE blocks. Spatial attention is used to capture long-range contextual information. Spatial attention is composed of strip pooling (Mixed Pooling Module).

3. The video shows the result difference between the SOTA and their model, however, it is unclear which portion of the model is the reference and which portion is the improvement.
   Our response:
   - Architecture: we concatenate the low level features to decoder in order to enhance the detail features. The model discussed in the original paper, the reconstruction of low features, such as texture and edges, is not that great. Thus, our model makes some improvement on those aspects.
   - Attention: We introduce the channel and spatial attention to our fusion model (Student model), so we are able to capture the correlation between channels and locations.
   - Low level feature knowledge distillation: The original paper uses the output of the Resblock layer, which means high level features, as the guidance of the student model. We believe that provides insufficient guidance of reproducing low level features. So we add in the low level feature guidance in our framework, so we are able to capture the low level features better.

Critical review from team 32:

1. The group could mention additional details of the dataset such as the number of image categories etc.
   Our response: We have already mentioned that in slides and talked about that in our presentation. Please refer to the report and slides. Each image was rendered with -1.5, -1, +0, +1, and +1.5 relative exposure values (EVs) by an emulation of the camera ISP processes. The dataset can be splitted into three sets: training set of 17,675 images, validation set of 750 images, and testing set of 5,905 images.

2. The quantitative and qualitative(visual) analysis should have been performed on more than one datasets if available.
   Our response: Due to our time constraints, we could only perform our training and testing on this dataset. In the future, we will try on more datasets.

3. Can the group explain the perceptual loss in more detail, especially how exactly ϕ is computed?
   Our response: ϕ is the output from VGG19, pre-trained on ImageNet. We compute the difference between restoration results and the ground truth at the layer of 2, 7, 12, 21, and 30 of VGG19**.**

4. The group should mention what pairs of exposure level images are being used as inputs to train the fusion and teacher models. For example whether images of exposure level 1.5 and 0 are used, or -0.5 and 0.5 are used etc. How does the training performance of the fusion model vary by changing the training data by using different combinations of exposure level images in training.
   Our Response: In the slides, we have mentioned that the models are trained using images that were randomly selected. In order to simulate the real world problems, we decided to use random selected images as our inputs. If our training data is only on low exposure images or high exposure images, our model would not be able to adapt different image input with different exposure levels.

5. Although the group has done quantitative analysis for attention based models, it would be better if the group provides qualitative evaluation of models using attention mechanism as well to better understand the effect of attention.
   Our Response: We will put the image before attention and after attention in the report**.**

Critical review from team 25:

1. The talk mentioned the model in “Distilling Image Dehazing with Heterogeneous Task Imitation” loses some feature representation and they are trying to do some improvements based on this model. However, I was not clear what improvements were made.
   Our response: This comment is similar to the one mentioned by group 21. So we will use the same answer. Here are the improvements we have made.
   - Architecture: we concatenate the low level features to decoder in order to enhance the detail features. The model discussed in the original paper, the reconstruction of low features, such as texture and edges, is not that great. Thus, our model makes some improvement on those aspects.
   - Attention: We introduce the channel and spatial attention to our fusion model (Student model), so we are able to capture the correlation between channels and locations.
   - Low level feature knowledge distillation: The original paper uses the output of the Resblock layer, which means high level features, as the guidance of the student model. We believe that provides insufficient guidance of reproducing low level features. We add in the low level feature guidance in our framework, so we are able to capture the low level features better.

2. The ratio of training dataset, validation dataset, testing dataset is about 24:1:8. I was curious why you choose to split data like this.
   Our response: This setting is from the original paper. In order to make a comparison between our results and theirs, we need to use the same setting.