DCDR-UNet: Deformable Convolution Based Detail Restoration via U-shape Network for Single Image HDR Reconstruction

Joonsoo Kim, Zhe Zhu, Tien Bau, Chenguang Liu
DMS Lab, Samsung Research America, USA
{joonsoo.k, zhe.zhu, t.bau, cheng.liu1}@samsung.com

1. Overview

As requested from workshop, we first include all the reviews we got from CVPR and our rebuttals for reviews. Also, we add our new responses for all the reviews. Our new responses include our modifications in the paper and newly conducted experimental results.

Beside this, we also provide more experimental results that we could not include in the main manuscript because of the limits of space. First, we visualize the feature maps of our network. Since our overexposed details are mainly reconstructed through the multiple DCRBs, we visualize the feature maps of the last DCRB. Also, as we mentioned in the main manuscript, we provide the tone mapped results of the same images that we show in the main manuscript so that normal exposed regions of all the images are more visible. Last, we provide more qualitative results (also tone mapped) so that how our network can restore the over-exposed objects or regions in different scenes.

2. Reviews and Rebuttal from CVPR and Our change

We got three reviews for this paper from CVPR conference. The rates were one weak accept and two weak rejects. Since we are required to put the reviews, we collect all the reviews and reorganize and show them here. For each review about paper weakness, we include our rebuttal and our new response in this paper. For each review, we also indicate which reviewer mentioned it. From now we denote reviewer 1, 2 and 3 as R1, R2 and R3.

2.1. Paper Strengths

1. The problem of this paper is important to address. (R1)
2. Literature review and evaluation seem fair. (R1)
3. Based on concerns about existing evaluation methods, the authors perform evaluation on datasets with known maximum peak luminance. Through this evaluation process, they have demonstrated that the proposed network produces qualitatively and quantitatively good results. (R2)
4. Sufficient ablation studies is conducted to validate the effectiveness of the proposed network architecture and the loss functions. (R1, R2, R3)
5. This paper is written concisely, effectively presenting its arguments. (R2)
6. The usefulness of deformable convolution filters is an interesting idea and has been explained well. (R3)
7. The test results on the chosen dataset indicate that their approach very convincingly outperforms other methods. (R1, R3)

2.2. Paper Weaknesses

1. We need to present experimental results on other HDR datasets. (R2, R3)

- Rebuttal :

As requested by two reviewers, we shortly conduct experiments on two additional datasets such as VDS dataset [6] and the dataset used in HDRUNET [2]. We included the results of our method and HDRUNET on both datasets in the rebuttal. In the rebuttal we showed that our method outperforms HDRUNET on both datasets. However, we get notified that there would be potential copyright issue on the dataset used in [2], so we decide that we do not include the results of the dataset here. However,we still include the experimental results we conducted on VDS dataset during rebuttal here, and the quantitative results are shown in Table 1. Even though our method achieves better than HDRUNET, the quantitative score on VDS dataset

Method	PSNR-L	PSNR-μ	HDR-VDP3
HDRUNET	22.31	18.04	5.35
Proposed	25.01	22.32	6.99

Table 1. Quantitative Scores on VDS dataset with incorrect experimental setting in the rebuttal

was much lower than the one on Cityscapes HDR dataset [3], which is the main dataset used in the main manuscript. The reviewer 2 and 3 said that our quantitative score is much lower than other existing methods reported in [1] and they gave me weak rejects in the final rate. However, we realize that our experiment setting for this dataset during rebuttal was not correct. The main problem was the data preprocessing. In rebuttal we normalized a HDR image by its maximum peak value for VDS dataset and it causes the images with very low peak luminance to be extremely brighter and the images with very high peak value to be extremely darker. It makes the details of these HDR images be too much attenuated. To solve this problem, we train and test our network on VDS dataset with new experimental setting again, which is described in next.

• Our new response :

Since most pixel values of HDR images (with hdr file extention) in VDS dataset are within [0, 1], we just do not apply normalization on HDR images in the new experimental setting. For 8 bit LDR images in VDS dataset, we still normalize them within [0, 1]. Then, we train our network using the pairs of the normalized LDR images and the original HDR images. Note that VDS dataset has 48 pairs of LDR and HDR images for a training set and another 48 pairs of them for a test set. We train our network on the training set and test it on the test set. We call the trained network on VDS dataset as ”Proposed (VDS)”.

Besides this, we also test our model trained on Cityscapes dataset (one used in our main manuscript) on VDS test set for cross dataset testing purpose. We call this trained network as ”Proposed (Cityscapes)”.

For the evaluation process and metrics for the experimental results on VDS dataset, we followed [1]. Because of insufficient time between main conference and this workshop, we train only our network on VDS dataset. For comparison with other existing methods, we borrowed the experimental results of prior works on VDS dataset from [1]. The quantitative results of our network and other existing methods are shown in Table 2. As we can see, our network trained on VDS dataset shows better results against most existing results. Note that the methods of [1, 4, 6, 7] are multiple exposure HDR reconstruction methods, which reconstruct multiple brackets LDR images first and combines them to generate an HDR image. In these methods, all the different exposure ground truth LDR images should be provided for training. Since all the tones and details of LDR images at all the different exposure levels are provided for training, we can say that these methods have more training resources than the methods that reconstruct an HDR image directly from an LDR image like our method. Despite this disadvantage, our method achieves better scores than most of the multiple exposure HDR reconstruction methods.

For our method (Cityscapes), we can see that it achieves much lower PSNR values than many other methods. That is because the inverse tones between LDR and HDR images in the Cityscapes dataset are very different from the inverse tones between LDR and HDR images in the VDS dataset. However, we find that the visual quality of this method still looks good. As we can see, even though the tones of HDR images in proposed (Cityscapes) are brighter than the tones of Ground Truth (GT) HDR images, it still restores the overexposed objects such as sky or window frames well. Also, the HDR-VDP2 Q score for Proposed (Cityscapes) is not bad as PSNR metric. That is because HDR-VDP2 Q score considers not only image tones but also many other factors to calculate the score.

2. Issues with referencing prior works. Since some of prior works use the perceptual loss, they should be cited. Also, since the entire network architecture is similar to HDRUNET, the similarity and the major difference between proposed method and HDRUNET should be addressed in details. (R3)

• Rebuttal :

In the rebuttal, we said that we would correct all the missing referencing the prior works in final manuscript if the paper is accepted.

• Our new response :

In the proposed method section in the main manuscript, we correctly reference the priors works that use the perceptual loss. Also, we describe what the similar parts and the major differences between our network and HDRUNET are in the proposed section of the main manuscript in details.

3. Lack of novelty: The idea of utilizing deformable convolution to expand the receptive fields is used in other low-level computer vision application. (R2)

• Rebuttal :

	PSNR RH's TMO		PSNR KK's TMO		HDR-VDP2	
Method	m	σ	m	σ	m	σ
DrTMO [4]	25.49	4.28	21.36	4.50	54.33	6.27
Deep chain HDRI [6]	30.86	3.36	24.54	3.50	56.36	4.35
Deep recursive HDR [7]	32.99	2.81	28.02	6.99	57.15	6.46
Mask HDRCNN [10]	22.56	2.68	18.23	3.53	53.51	4.76
SingleHDR [8]	30.89	3.27	28.00	4.11	56.97	6.15
CEVR [1]	34.67	3.50	30.04	4.45	59.00	5.78
Proposed (Cityscapes)	24.22	3.09	21.00	4.40	56.99	5.85
Proposed (VDS)	33.61	4.66	30.38	6.40	61.11	7.14

Table 2. Quantitative Scores on VDS dataset with correct experimental setting (Red : the best, Blue : the second, Green : the third). RH's TMO and KK's TMO represent tone mapping functions used in [9] and [5]

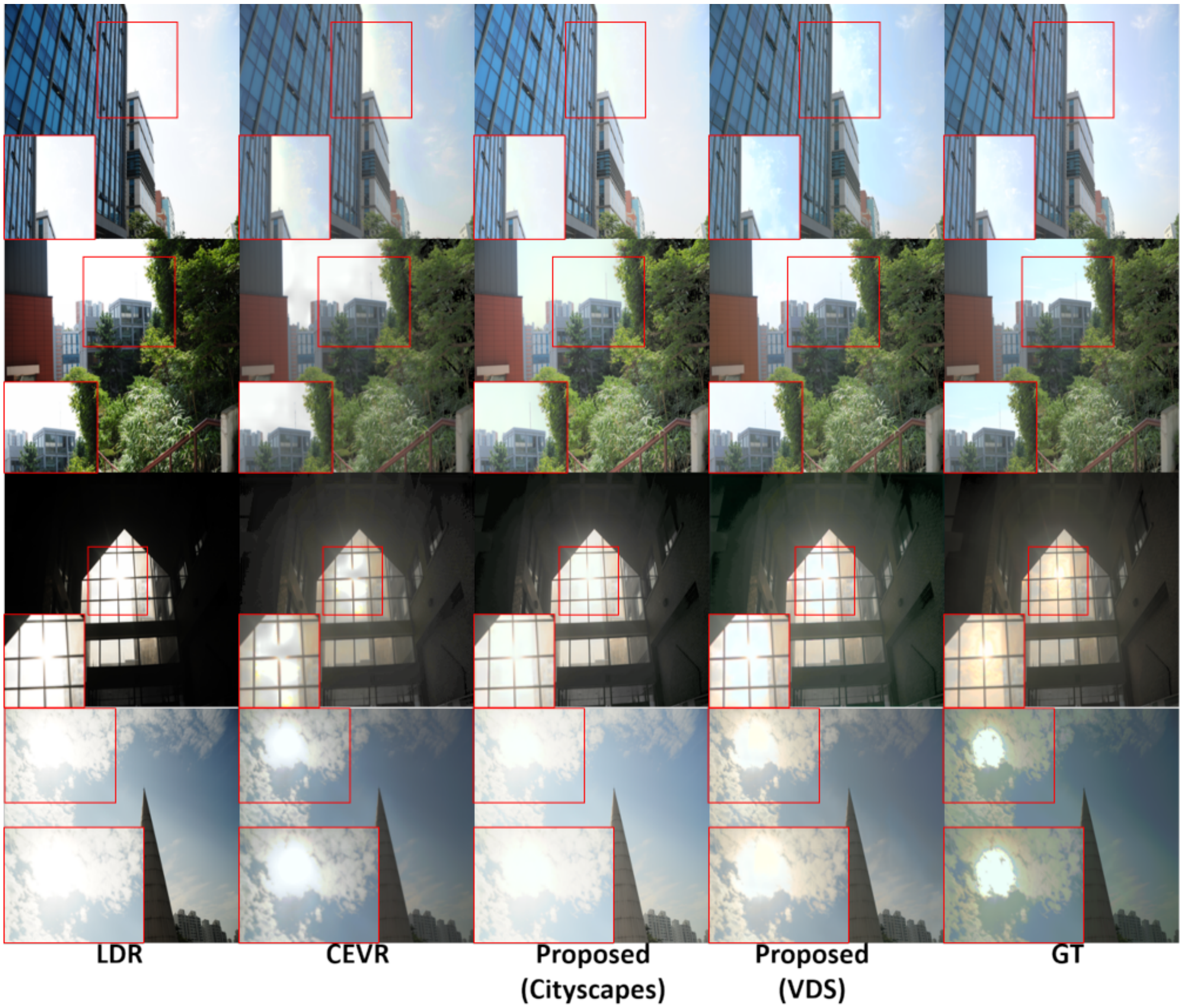

Figure 1. Qualitative Results on VDS dataset. These results are tone mapped using the tone mapping method from [5]

Since we describe that utilizing deformable convolution in single HDR reconstruction is not only to expand the receptive fields but also to make the receptive fields learnable and it is critical to restore partially overexposed objects such as overexposed power lines or overexposed tree branches, in the main manuscript, we do not address this problem in the rebuttal again.

• Our new response :

We more emphasize that utilizing DCRB, which uses deformable convolution, can help the receptive field of each pixel be learnable in the main manuscript.

4. Why the designed network produces strong effects in detail is not clear. We need the detailed explanation for this. (R2)

• Rebuttal :

We described the main design points of DCRB during rebuttal. Which are (1) putting deformable convolution in the residual path of DCRB while putting offset estimation on the input path and (2) training it with pixel loss + perceptual loss.

(1) Offsets in a DCRB can be accurately estimated based on the restored details from previous DCRBs while deformable convolution on the residual path focuses on restoring additional details. This DCRB design helps to restore the details progressively and more accurately through multiple DCRBs.

(2) In DCRB the offset is estimated through minimizing the training loss. With pixel loss only, the offset is not estimated to indicate the non-overexposed pixels of the partially overexposed object. Instead, it is estimated to indicate similar tone pixels. With additional perceptual loss, the offsets on the overexposed pixel of the object are estimated toward the non-overexposed pixels of the same object. It forces the learned receptive field to include the restoring object compactly (as shown in power line example in Figure 1 in the main manuscript).

• Our new response : To understand better how the deformable convolution works to produce the details, we added the additional figure (Figure 2) and description in the introduction of the main manuscript.

5. There are typos in the paper. (R1, R2, R3)

• Rebuttal :

In the rebuttal, we said that we would correct all the typos in the final manuscript if the paper is accepted.

• Our new response :

We fixed all the typos mentioned in the reviews.

6. Use HDR-VDP2 for evaluation metric. (R2)

• Rebuttal :

Since we used HDR-VDP3, which is the updated version of HDR-VDP2, in the main manuscript, we did not address this problem during rebuttal.

• Our new response :

We found that HDR-VDP2 is still dominant metric compared to HDR-VDP3 in the existing single HDR reconstruction methods. To make our experimental results comparable to the experimental results reported in the other singe HDR reconstruction papers, we recalculate HDR-VDP2 Q scores and put them into the main manuscript.

3. Visualization

To understand how DCRB restores the large overexposed region as well as the partially overexposed small object with sparse information, we visualize the feature maps of DCRB when we input I_{ldr} into our network. As shown in the fourth column images in Figure 3, only our method restores both the complete small tower and natural sky region. Figure 2 shows how these objects are restored. Even though the top of the tower looks broken in the input image (LDR), we can see that the broken part is activated in F_1^{DCRB}, the first feature map of the last DCRB in our network. It means that the broken part of the tower is restored through multiple DCRBs. Also, F_2^{DCRB}, the second feature map of the last DCRB, shows that the overexposed sky region is activated here. It means that the color and texture of the sky region is also restored through multiple DCRBs.

Besides that, we also visualize the outputs of $Net_T(I_{ldr})$ and $Net_R(I_{ldr})$ to prove that our restoration network mainly restores the overexposed regions while our tone network focuses tone matching on normal exposure region between LDR and HDR images. As we can see in the third column images in Figure 2, $Net_R(I_{ldr})$ restores the details of the overexposed region only (almost no restored information on the normal exposed region). $Net_T(I_{ldr})$ mainly restores the tones of normal exposure region but no details in the overexposed region.

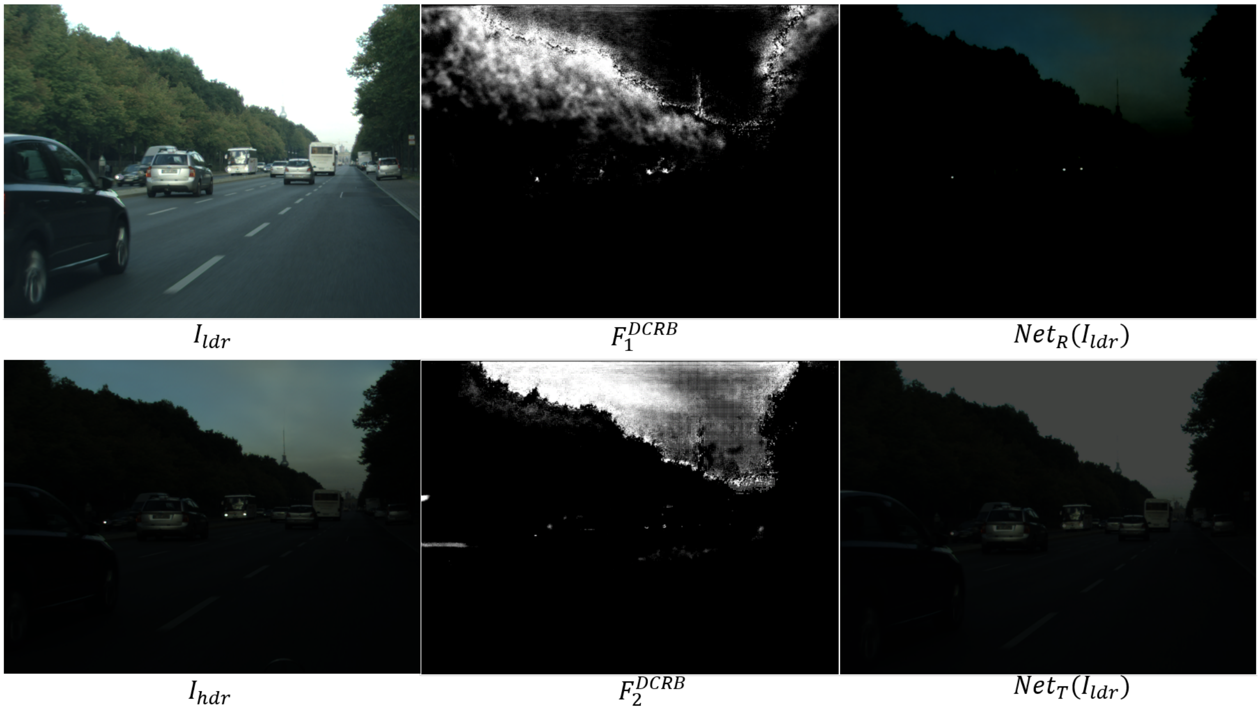

Figure 2. Visualization of feature maps. F_1^{DCRB} and F_2^{DCRB} is the 1^{st} and 2^{nd} feature maps of the last DCRB in our network. The broken top of a small tower is restored in F_1^{DCRB} and the overexposed sky is restored in F_2^{DCRB}.

4. Additional Qualitative Experiments

Through Figure 3, 4, 5, we provide more qualitative results. Different from the qualitative results in the main manuscript, all the images are tone mapped. For the tone mapping, we use the following tone mapping algorithm, which is used in [2].

$$I_{hdr-n-tone} = \frac{log(1 + \mu \times Tanh(I_{hdr-n}))}{log(1 + \mu)} \tag{1}$$

where I_{hdr-n} is the HDR image normalized by $2^{16} - 1$ so that it can be within [0,1]. $I_{hdr-n-tone}$ is the tone mapped HDR image from I_{hdr-n}. For μ, we use $\mu = 10$ empirically so that the normal exposed regions of all the images are more visible and the tone mapping does not attenuate the restored details much. If we use the higher value such as $\mu = 5000$, which is used in [2], it removes most of the restored details by making the bright regions be much brighter (since $\mu = 10$ is relatively small, some normal exposed regions still look dark).

For better visual comparisons, we categorize the overexposed objects/regions into 6 categories (power line (thin line), small structure, car, sky/clouds, tree, building). Note that we mainly focus on the restored overexposed regions in these comparisons.

As we can see in Figure 3, our method can restore the entire power lines, which look disconnected in LDR images, much better than all the existing methods. Even though some of the existing methods also restore some parts of the power lines, their results still look disconnected. However, the most lines are reconnected through our method.

For small structures, our method also restores the partially overexposed parts of them. As shown in the fourth to sixth column in Figure 3, most of them are naturally reconstructed in our method compared to most existing methods.

For the overexposed cars shown in Figure 4, our method restores the smooth surfaces of them closer to ground truth (GT) than the existing methods that restore the surface non-smoothly.

For the overexposed sky/clouds shown in Figure 4, the existing methods with large network generally reconstruct the sky naturally while the methods with small network reconstruct sky unnaturally. This is because of small fixed receptive fields of them. However, our method reconstructs the sky/clouds very naturally even though our network is relatively small. Note that if there is no information of clouds in the overexposed sky in the LDR image, our method mainly reconstructs sky only. Therefore, our reconstructed sky looks different from the one from ground truth, but it looks natural.

For the partially overexposed trees shown in Figure 5, some parts of leaves are overexposed, which look them to be disconnected from the trees. In our reconstructed HDR image, the leaves are reconnected to the trees. However, most existing methods fail to reconnect them to the trees.

As shown in Figure 5, the surfaces of buildings are partially overexposed. While most exiting methods reconstruct the surface textures unnaturally or the surface color differently from ground truth, our method reconstructs the natural surface textures and colors closer to ground truth.

Through these additional visual comparisons, we prove that our method can reconstruct the details in the overexposed region more naturally than the existing methods for many cases.

References

[1] Su-Kai Chen, Hung-Lin Yen, Yu-Lun Liu, Min-Hung Chen, Hou-Ning Hu, Wen-Hsiao Peng, and Yen-Yu Lin. Learning continuous exposure value representations for single-image hdr reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12990–13000, 2023. 2, 3

[2] Xiangyu Chen, Yihao Liu, Zhengwen Zhang, Yu Qiao, and Chao Dong. Hdrunet: Single image hdr reconstruction with denoising and dequantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 354–363, 2021. 1, 5

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[4] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Transactions on Graphics (Proc. of SIGGRAPH ASIA 2017)*, 36(6), 2017. 2, 3

[5] Min Kim and Jan Kautz. Consistent tone reproduction. *Proceedings of the 10th IASTED International Conference on Computer Graphics and Imaging, CGIM 2008*, 2008. 3

[6] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep chain hdri: Reconstructing a high dynamic range image from a single low dynamic range image. *IEEE Access*, 6:49913–49924, 2018. 1, 2, 3

[7] 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. 2, 3

[8] Y. Liu, W. Lai, Y. Chen, Y. Kao, M. Yang, Y. Chuang, and J. Huang. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[9] Erik Reinhard, Michael Stark, Peter Shirley, , and James Ferwerda. Photographic tone reproduction for digital images. *ACM Transactions on Graphics*, 2(69):661–670. 3

[10] Marcel Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image hdr reconstruction using a cnn with masked features and perceptual loss. *ACM Transactions on Graphics*, 39, 2020. 3

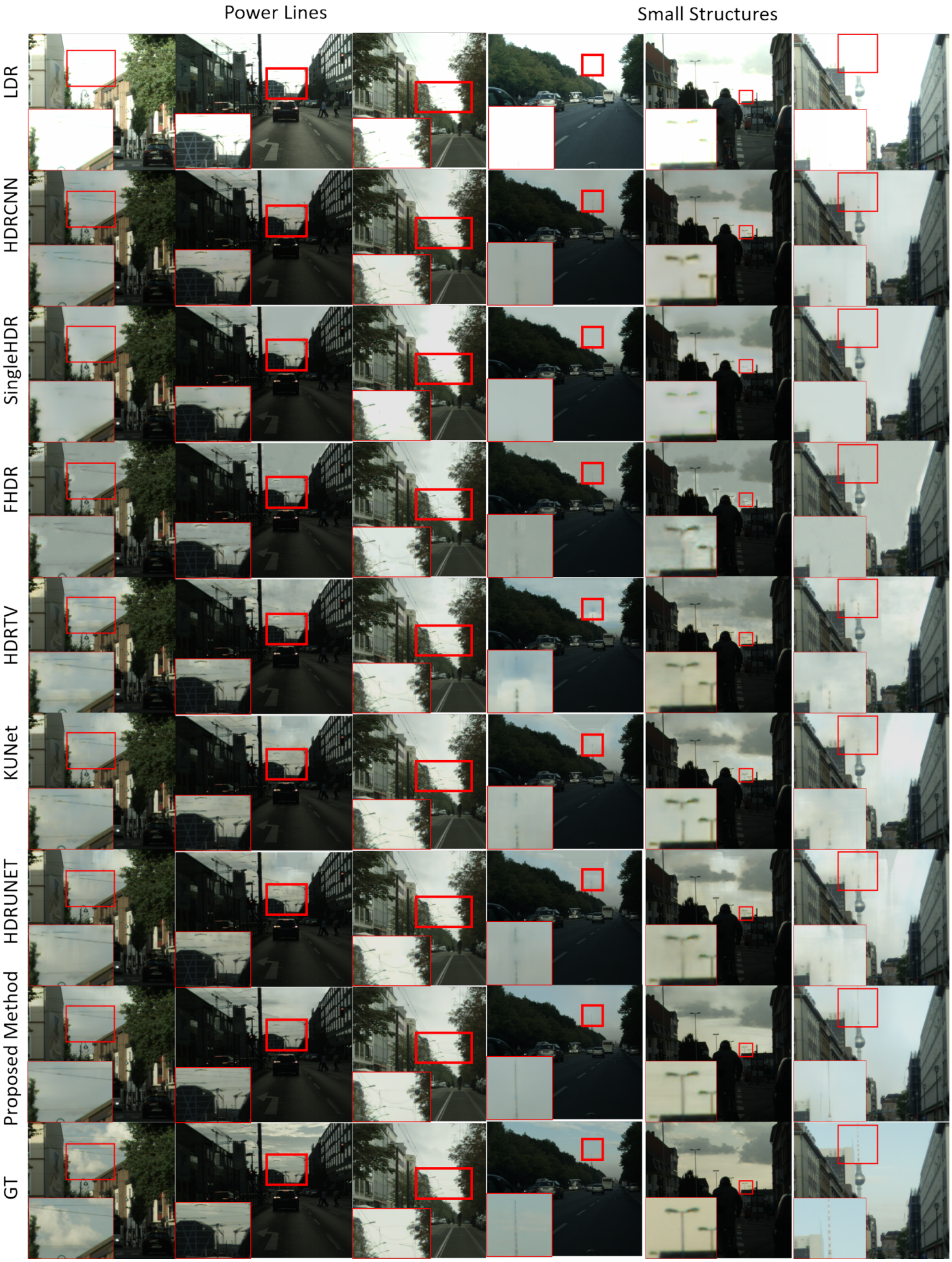

Figure 3. Qualitative Comparison with tone mapping. These difference are better visible in pdf version with zoom-in.

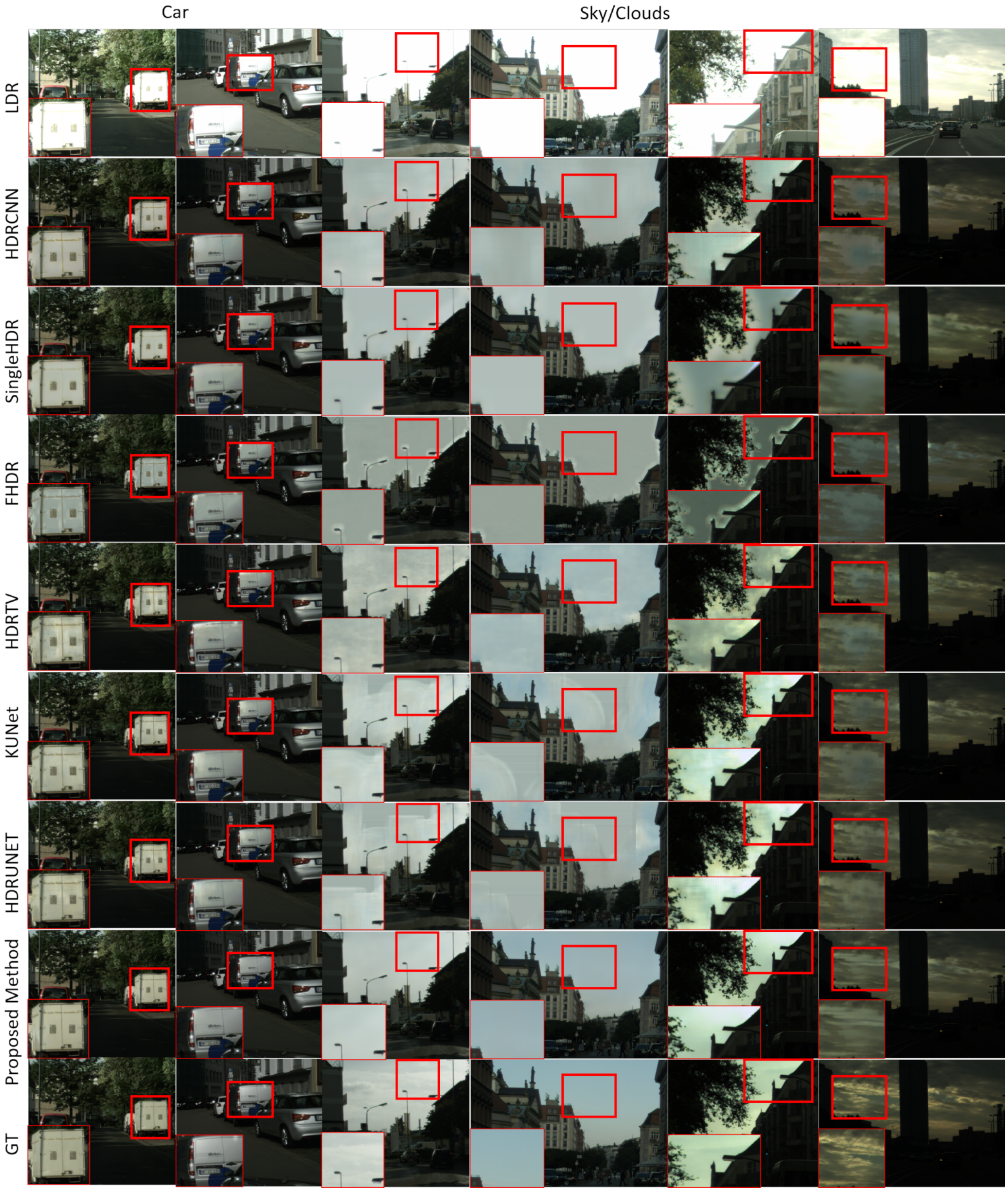

Figure 4. Additional Qualitative Comparison with tone mapping. These difference are better visible in pdf version with zoom-in.

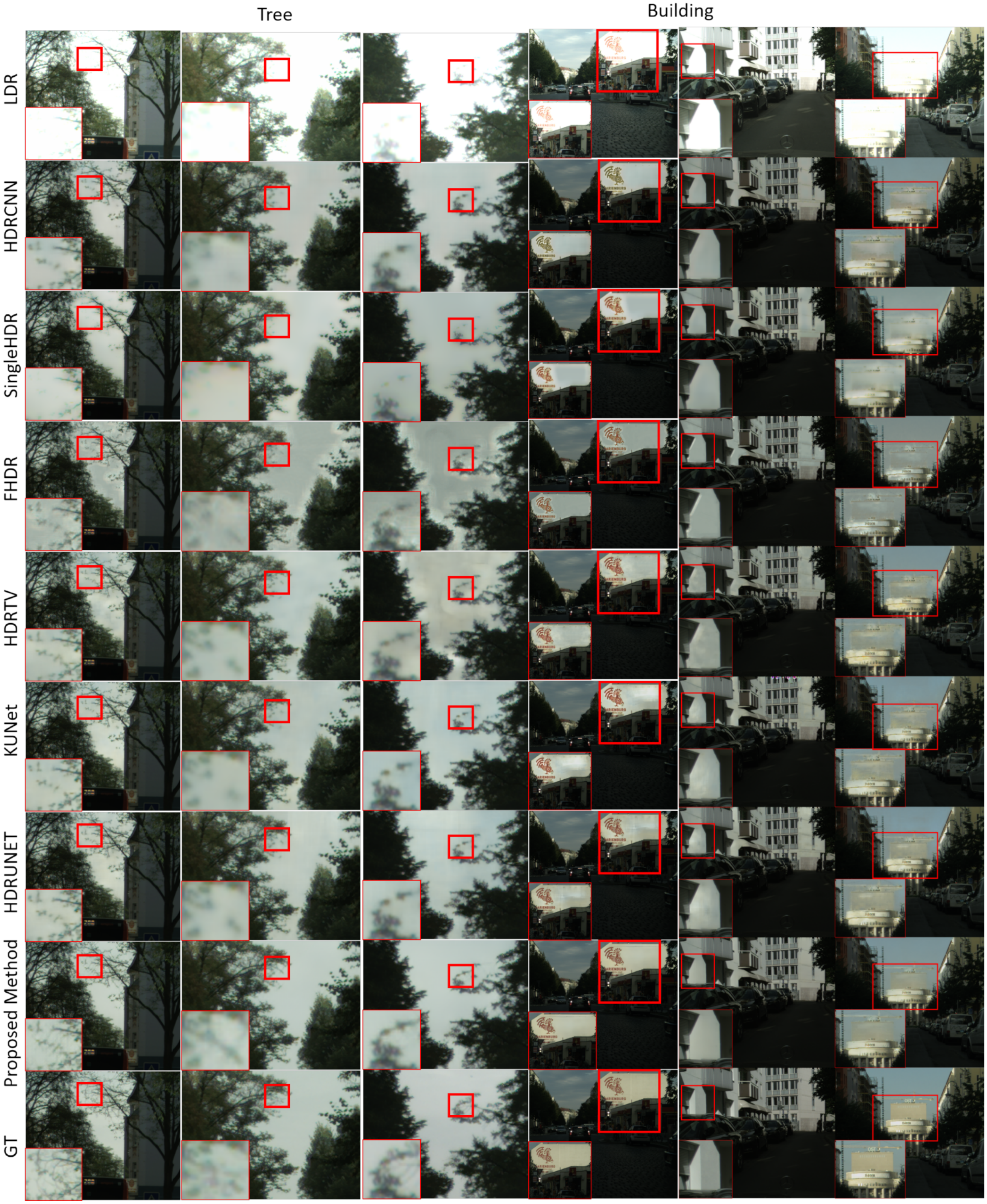

Figure 5. Additional Qualitative Comparison with tone mapping. These difference are better visible in pdf version with zoom-in.