Intrinsic Single-Image HDR Reconstruction
Supplementary Material

Sebastian Dille*, Chris Careaga*, and Yağız Aksoy

Simon Fraser University
Burnaby, BC V5A 1S6, Canada
sdille@sfu.ca

In this supplementary document, we present (i) a detailed qualitative analysis of our approach as an extension of Section 5.2 of the main paper, (ii) an extension of our quantitative analysis in Section 5.1 of the main paper, (iii) an ablation study regarding the different parts of our pipeline, and (iv) an in-depth description of the training process including a runtime analysis of our final approach.

A Qualitative Analysis

We provide an extensive qualitative evaluation against state-of-the-art HDR reconstruction methods. We focus our evaluation on the SIHDR Benchmark dataset [9] as it consists of everyday scenes at high resolutions.

Figure 1 shows our method compared to 7 prior works. The images provided specifically highlight the ability of our model to recover over-saturated and clipped colors. Many of the competing methods often fail to recover clipped image information, resulting in desaturated colors. This can be seen in the results of ExpandNet [15], HDRUNet [3], DrTMO [6], HDRCNN [5] and MaskHDR [19] particularly in images with over-exposed sky regions like the [forest] and [sunset] scenes. Although the method of Liu *et al.* [13], SingleHDR, is able to recover clipped regions, the colors are often over-saturated giving the resulting HDR image an unrealistic appearance. This can be seen in the sky of the [sunset] scene, and on the door in the [deck] scene. Our method can recover accurate color information, even when large regions of the image are lost, which can especially be seen in our accurate sky regions. Furthermore, ours is the only method that is able to recover the proper saturation of the background buildings in the [man] scene. We attribute our improved color reconstruction to our intrinsic formulation. Figure 3 shows our estimated components for the same scenes in Figure 1. Although the LDR albedo contains corrupted values from over-exposed regions, our estimated HDR albedo accurately in-paints the lost information giving our final refined result a realistic color appearance.

We show the same set of methods on 4 more scenes in Figure 2. These scenes highlight our method's ability to recover high-resolution details in over/under-exposed regions. Prior works struggle with predicting accurate details around edges where over- and under-exposed pixels meet. This can be seen in the [forest]

(*) denotes equal contribution.

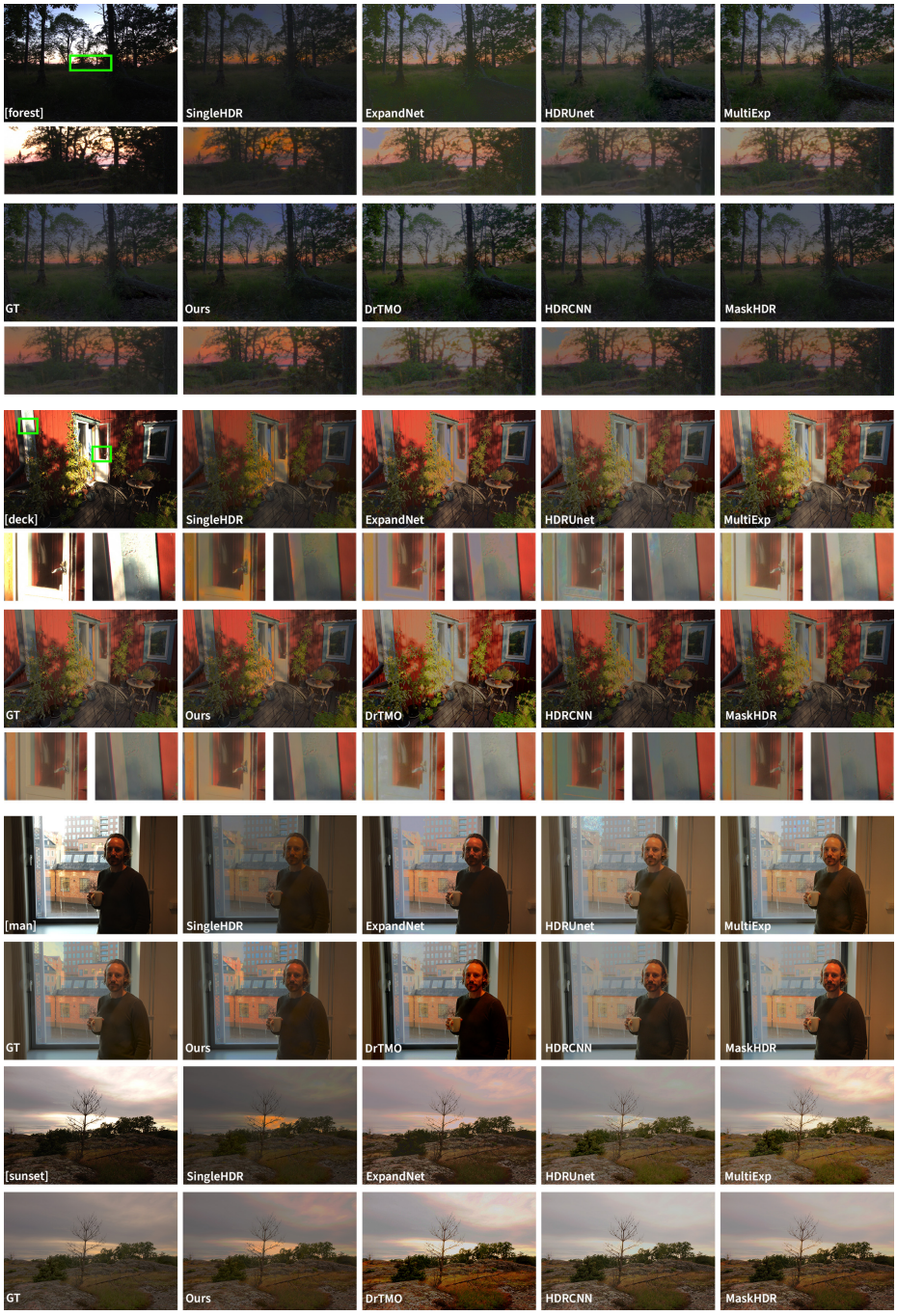

Fig. 1: We show qualitative results on the SiHDR benchmark [9] with a focus on color reconstruction, tone-mapped for visualization via Photoshop. We refer to Section A for an in-depth discussion.

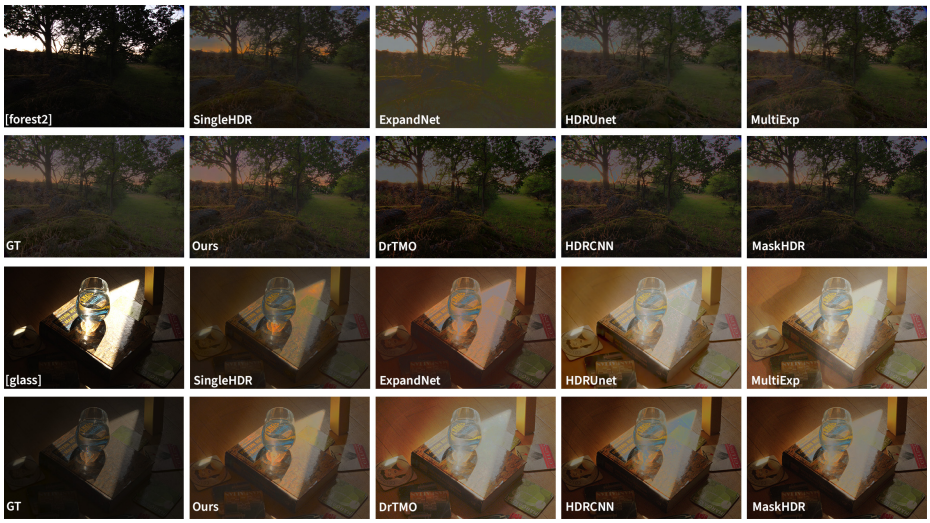

Fig. 2: We show qualitative results on the SiHDR benchmark [9] with a focus on detail recovery, tone-mapped for visualization via Photoshop. We refer to Section A for an in-depth discussion.

Fig. 3: We show examples of the reconstructed HDR Albedo and Shading for images from SI-HDR [9], tone-mapped for visualization via Photoshop. We refer to Section A for an in-depth discussion.

and [sunset] scenes. Ours is the only method able to recover both accurate colors and details on the cover of the textbook in the [glass] scene. All other methods either distort the colors or leave the cover over-exposed. We attribute our detail recovery to our HDR shading formulation. Figure 4 shows the predicted intrinsic components for the scenes in Figure 2. We can see our shading model is able to reconstruct accurate details and extended dynamic range for regions lost due to over-exposure. This comes from the simplified task for our HDR shading network where the task only requires generating a wider dynamic range without accounting for color, which is handled in our HDR albedo formulation, and takes the LDR shading as input which shows a high correlation with the scene geometry compared to the LDR input image.

We provide additional qualitative comparisons against the two prior works with the strongest quantitative performance, SingleHDR [13] and HDRUNet [3]. Figure 5 shows our model's color reconstruction capabilities against these methods. We observe that our method is able to recover underlying color information in diverse scenarios. This can be seen in the [apple] scene on both the leaves and the apple itself. SingleHDR alters the color of the apple and adds a yellow tinge to the over-saturated leaf regions, while HDRUNet is not able to recover the lost color information at all. Similar behavior is seen in the [toys] scene. While our method recovers accurate colors in the sky and on the deck, SingleHDR under-saturates the sky and over-saturates the deck while HDRUNet introduces artifacts to both. In the [drawers] scene, although our method does predict a more saturated orange than the ground truth, the color is uniform and realistic. SingleHDR predicts distorted colors and adds a green tint overall. HDRUNet is not able to recover the underlying color and leaves many over-exposed regions. Finally, unlike the other methods, our approach is able to generate an accurate roof color for the background of the [candles] scene. Figure 7 shows our predicted intrinsic components for some of these scenes. The colors we see recovered in the final result come from the estimated HDR albedo. This can be observed both on the apple and on the wooden deck.

Figure 6 shows our model's detail reconstruction against the same two methods. In the [sunglasses] scene, our method is able to predict small details on the clouds in the reflection of the lenses. The result from SingleHDR is missing this detail, and HDRUNet adds incorrect texture and color to the clouds. In the [window] scenes our method properly reasons about the texture and color of the grass in the background while SingleHDR produces yellow artifacts and HDRUNet flattens the over-exposed region. Again, we can attribute our superior performance to our HDR shading network. The components for two of these scenes are shown in Figure 7. Our method recovers fine shading details on the over-exposed regions of the children's faces and estimates accurate shading on the contours of the bust.

It is important to note that the qualitative observations made with respect to our model do not necessarily correlate with the quantitative scores our model achieves. We observe a particular discrepancy in the PSNR scores of our model when compared to prior works. We provide quantitative scores alongside the

Table 1: Quantitative results against state-of-the-art on the challenging SI-HDR [9] benchmark.

method	PSNR	VSI	HDR-VDP3	PSNR-H
OURS	36.62	98.27	**8.96**	32.63
DrTMO [6]	33.58	96.73	8.27	28.50
HDR-CNN [5]	35.91	98.17	8.39	31.87
ExpandNet [15]	36.01	97.65	8.67	32.53
Single-HDR [13]	35.68	**98.30**	8.79	31.34
Mask-HDR [19]	36.72	98.22	8.25	**33.28**
HDRUNet [3]	**36.92**	98.16	8.82	31.58
Multi-Exp Gen. [11]	35.36	98.01	8.64	31.23
Lightweight [8]	34.14	96.95	8.26	29.32

scenes shown in Figures 5 and 6, and Figure 8 in the main paper. The [toys] scene shows this discrepancy well, the result from HDRUNet with artifacts in the over-exposed regions yields a PSNR of 36.4. This is nearly two points higher than the score yielded by our method despite our accurate reconstruction. The VDP scores for this image on the other hand seem to more accurately correlate with the subjective quality of each result. This aligns with the findings of Hanji *et al*. [9] and is reasonable given that HDR-VDP specifically models human perception of HDR imagery [14]. One notable exception to this trend is the [bust] scene in Figure 8 where HDRUNet achieves the highest VDP despite not recovering any over-exposed information. Despite these exceptions, we find that VDP is the most accurate quantitative indicator of subjective quality and therefore focus on this metric in our quantitative evaluation.

B Extended Quantitative Analysis

We extend our main evaluation in Table 1 with the VSI [21] metric to gain additional insights into our model's performance. Since VSI is not designed for high-dynamic range values, we convert the reconstructed images together with the corresponding ground truth to PU21-space [1] as recommended by Hanji *et al*. [9]. We additionally report the accuracy in the highlight region indicated by PSNR-H with the same input encoding to focus on our reconstruction intent. For this purpose, we restrict the PSNR metric on pixels with values $I > 0.8$ in the input LDR image. As detailed in Section A and visualized in Figures 5 and 6, the PSNR metric does not correlate well with perceived visual quality despite the PU21 color space. We focus our discussion in the following on HDR-VDP3, VSI, and PSNR-H. For completeness following the evaluation settings in prior work, we add PSNR to the provided Tables 1–5.

We show the additional results for the SI-HDR [9] benchmark in Table 1. SI-HDR [9] is a challenging dataset. For one, it is a zero-shot setting for all reported methods in our evaluation i.e. the dataset does not overlap with the

Table 2: Quantitative results against state-of-the-art on the RAISE [4] test split. We indicate an overlap with the training distribution of the full system with orange.

method	PSNR	VSI	HDR-VDP3	PSNR-H
OURS	32.38	97.95	8.89	32.21
DrTMO [6]	32.04	97.85	8.64	32.23
HDR-CNN [5]	31.08	97.61	8.47	30.44
ExpandNet [15]	31.13	97.62	8.27	30.14
Single-HDR [13]	**33.03**	**98.16**	**8.95**	**33.23**
Mask-HDR [19]	31.23	97.67	8.44	30.69
HDRUNet [3]	30.45	96.64	8.04	28.95
Multi-Exp Gen. [11]	31.78	97.94	8.66	31.67
Lightweight [8]	29.07	95.88	7.68	28.02

Table 3: Quantitative results against state-of-the-art on the HDR-Eye [17] test split. We indicate an overlap with the training distribution of the full system with orange.

method	PSNR	VSI	HDR-VDP3	PSNR-H
OURS	33.65	97.42	**8.96**	30.28
DrTMO [6]	32.14	97.04	8.72	28.27
HDR-CNN [5]	31.65	96.56	8.61	27.00
ExpandNet [15]	31.83	96.66	8.49	28.26
Single-HDR [13]	**34.07**	**97.59**	**8.96**	**31.13**
Mask-HDR [19]	31.84	96.67	8.60	27.32
HDRUNet [3]	31.80	96.15	8.46	26.94
Multi-Exp Gen. [11]	33.27	97.54	8.87	30.59
Lightweight [8]	29.92	94.78	8.20	25.12

training distribution of any of the methods. Good reconstruction results on this benchmark thus require the model to generalize to novel distributions and are the only meaningful indicators for real-world performance on images in-the-wild. Second, its high resolution requires the ability to reconstruct high-frequency changes in luminance accurately. Neural networks are often limited by their architecture in their capacity to account for small details, limiting the faithful recovery of intricate structures. One example is the [roof] scene in Figure 2 showing a tree with the sky visible through the leaves. Our method is successfully able to reconstruct the colors of the sky while preserving the structure of the leaves in contrast to many competitors. This behavior translates to the numerical performance. While the results for the individual metrics vary, our method ranks first overall with a lead in the HDR-VDP3 as the only metric directly targeted for HDR content and close second places in both VSI and PSNR-H. Especially the latter indicates that our method can truthfully recover missing color and details in the highlight regions of the image.

As additional comparisons, we include the performance on the test sets from RAISE [4], HDR-Eye [17], HDR-Synth [13] and HDR-Real [13] as provided by

Fig. 4: We show examples of the reconstructed HDR Albedo and Shading for images from SI-HDR [9], tone-mapped for visualization via Photoshop. We refer to Section A for an in-depth discussion.

Table 4: Quantitative results against state-of-the-art on the HDR-Synth [13] test split. We indicate an overlap with the training distribution of the full system with orange.

method	PSNR	VSI	HDR-VDP3	PSNR-H
OURS	32.36	96.38	8.21	29.53
DrTMO [6]	32.30	96.39	8.27	29.38
HDR-CNN [5]	31.58	95.82	8.03	28.66
ExpandNet [15]	31.10	95.91	7.69	27.94
Single-HDR [13]	**33.77**	**97.07**	**8.51**	**31.82**
Mask-HDR [19]	32.00	96.18	8.14	29.31
HDRUNet [3]	30.81	94.97	7.61	28.08
Multi-Exp Gen. [11]	32.55	96.34	8.16	30.14
Lightweight [8]	29.20	94.27	7.31	26.14

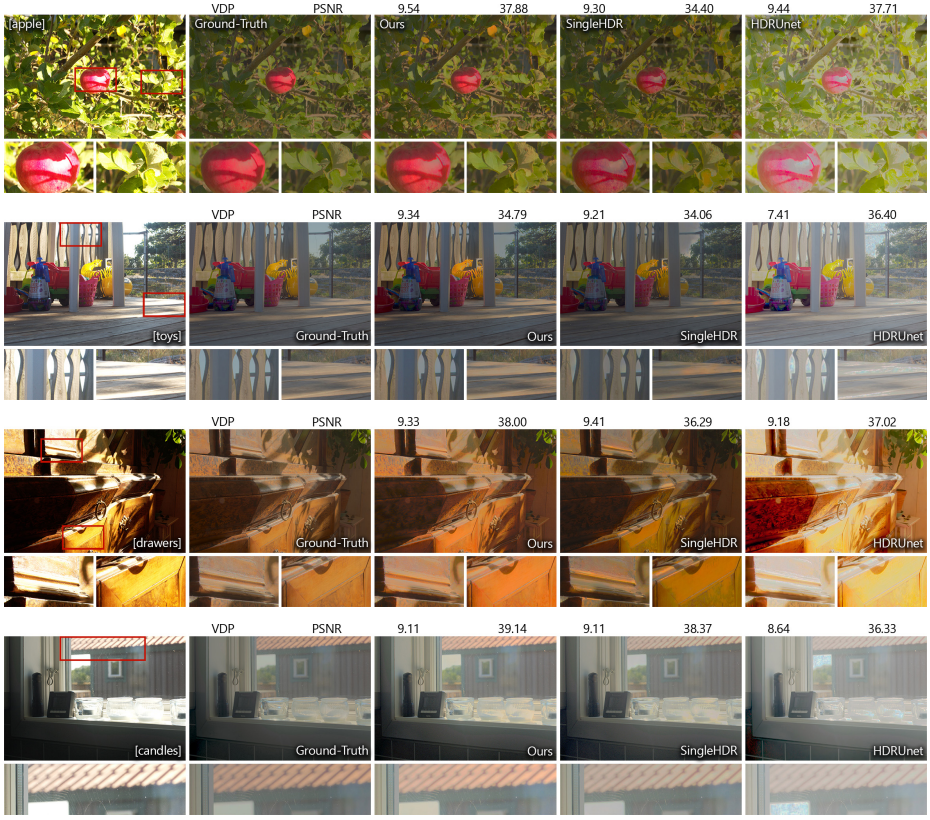

Fig. 5: We show qualitative results on the SiHDR benchmark [9] in comparison to SingleHDR [13] and HDRUnet [3] with a focus on color reconstruction, tone-mapped for visualization via Photoshop. We refer to Section A for an in-depth discussion.

Liu *et al.* [13] in Tables 2, 3, 4, and 5. Note that these evaluations are still zero-shot settings for our method. Other baselines in contrast use the respective training sets to train their final system which can then adjust to the image distribution in training. We indicate an overlap with the training distribution of the full system in orange in the tables provided. Our method achieves competitive results on all datasets, outranking the baselines on HDR-VDP3 [14] and coming close second in most other cases. The results become clearer if we eliminate the baselines with a distribution overlap for each dataset. In this zero-shot setting, our approach ranks first in all benchmarks.

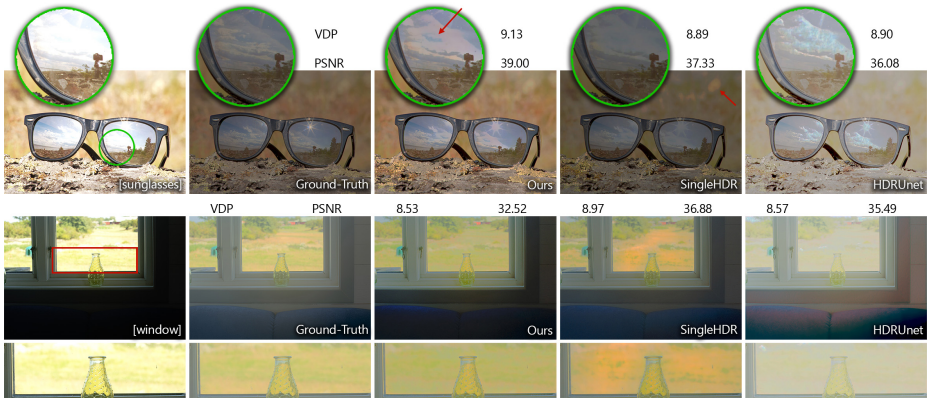

Fig. 6: We show qualitative results on the SiHDR benchmark [9] in comparison to SingleHDR [13] and HDRUnet [3] with a focus on detail recovery, tone-mapped for visualization via Photoshop. We refer to Section A for an in-depth discussion.

C Ablation Study

We analyze the contribution of the individual parts of our pipeline with a set of ablation studies. We start with the full proposed method as described in Section 4 of our paper and remove different parts individually. For all sets, we train the reconstruction networks for 150,000 iterations each and the refinement network for 75,000 iterations. In line with our main evaluation, the reported numbers are evaluated on the SI-HDR [9] benchmark dataset.

Intrinsic Reconstruction We show the effect of our losses on the reconstruction performance in Table 6. Removing the multi-scale gradient loss $\mathcal{L}_{MSG}$ has a small effect on HDR-VDP3 and PSNR but a comparably large influence on PSNR-H. This indicates that our method benefits from gradient-based supervision to accurately follow local changes in the shading, especially in the clipped image areas.

Secondly, we remove the reconstruction-based losses on the inferred modalities, leaving only the direct supervision against ground truth albedo and shading for the individual network, respectively. The significant drop in all metrics shows that the additional supervision is a necessary component to constrain the training and encourages faithful reconstruction via Eq. 1.

Refinement We show the effect of our losses and the different inputs on the performance of the refinement network in Table 7. For the gradient-based loss, we see a similar behavior to the reconstruction networks. This is expected, since both modules target similar goals for the reconstruction.

Interestingly, the model shows different behavior between removing the individual intrinsic components or the direct reconstructed $\hat{J}_{HDR}$. While all three inputs are shown to be beneficial for the training, removing $\hat{J}_{HDR}$ leads to a

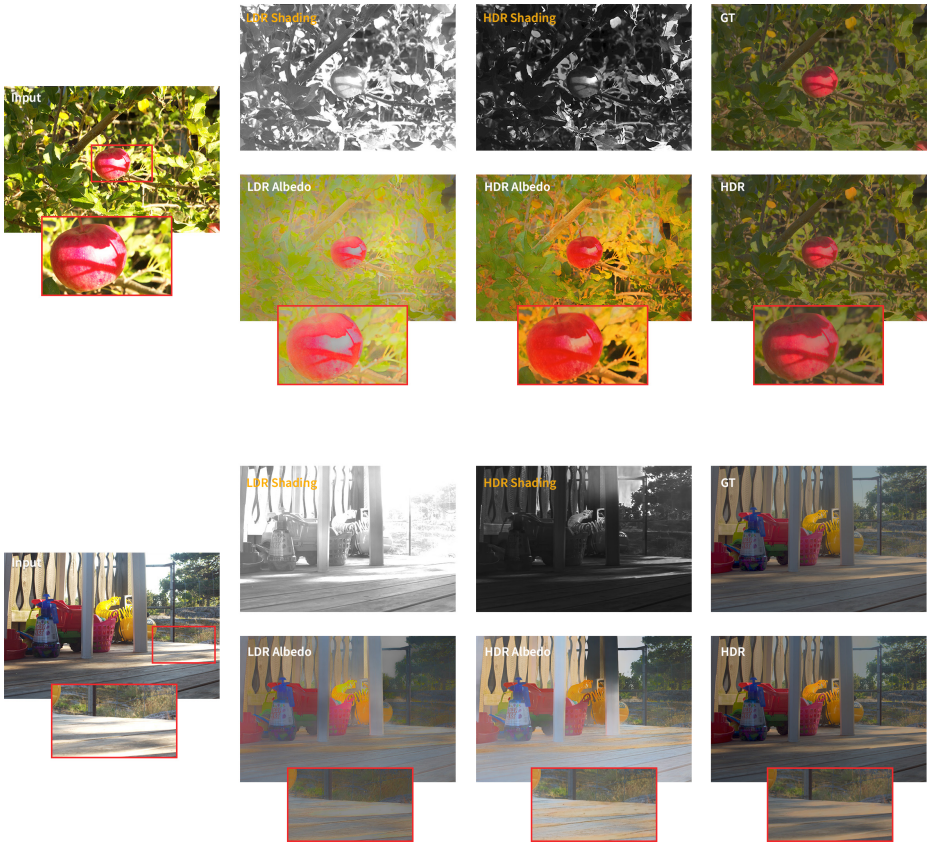

Fig. 7: We show examples of the reconstructed HDR Albedo and Shading for images from SI-HDR [9], tone-mapped for visualization via Photoshop. We refer to Section A for an in-depth discussion.

larger decrease in the metrics. We argue that inferring the connection between estimated albedo and estimated inverse shading poses a larger challenge for the model while the information provided by $\hat{J}_{HDR}$ is directly useable for inference.

Lastly, we do not see a large decrease in performance by training solely on Hypersim [18]. While this effect can be partially attributed to the large diversity of indoor scenes within the data, it cannot fully explain the performance on SI-HDR [9] which consists of mostly outdoor scenes. We interpret this as an indication that our intrinsic formulation introduces a higher degree of abstraction and thus increases the model's ability to generalize.

D Training details

We train our shading and albedo reconstruction methods with intrinsic ground truth from Hypersim [18] and the MultiIllum [16] dataset. We normalize the

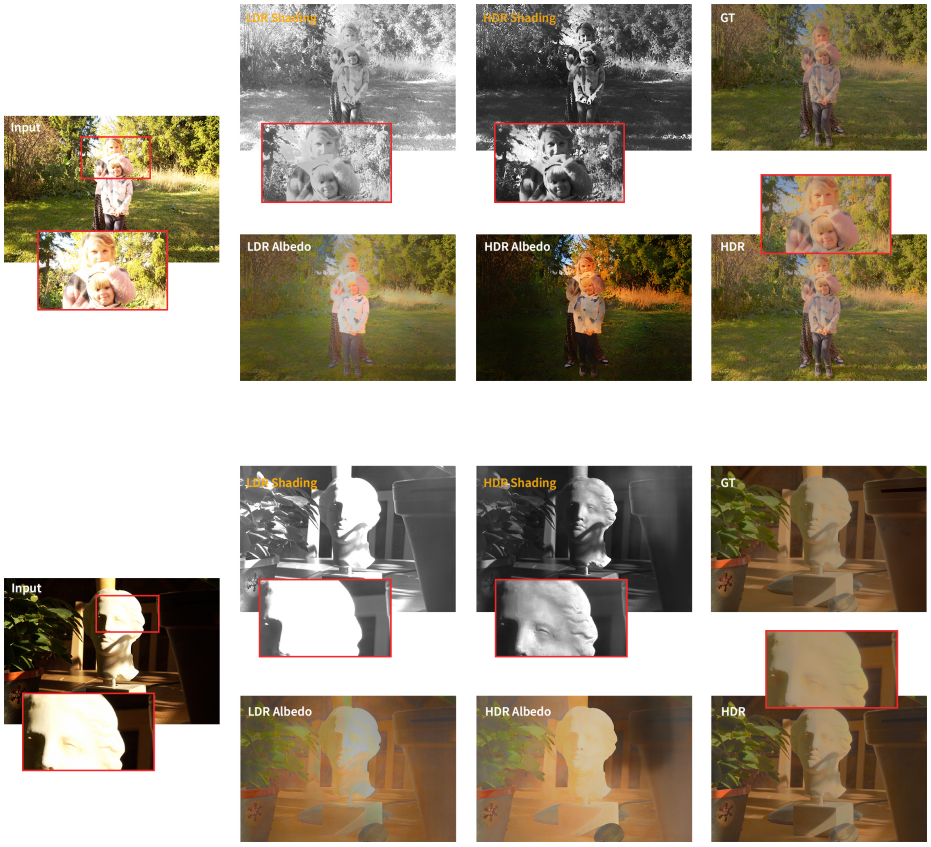

Fig. 8: We show examples of the reconstructed HDR Albedo and Shading for images from SI-HDR [9], tone-mapped for visualization via Photoshop. We refer to Section A for an in-depth discussion.

input HDR RGB images to a mean of 0.5 and apply a random exposure scale $e = 2^t, \quad t \in [-3..3]$. Since the intrinsic formulation is scale-invariant, we can derive the normalized and exposed intrinsics by dividing the modified RGB image by the provided albedo, applying the brightness change only on the shading. We then clip the RGB images before additionally applying random vertical and horizontal flipping and random cropping as augmentation, and scaling the resulting crop to the final training resolution of 512px as input to the decomposition network.

Note that due to the scale-invariance, we cannot rely on the decomposition network to retrieve A_L and S_L with the same scale as A_H and S_H. To stabilize the training process, we therefore match the scale of A_H to that of the LDR

version via least-squares before inference, yielding A_H^* as ground truth.

$$A_H^* = cA_H, \quad c = \arg\min_x \sum_i (xA_H - A_L)^2 \tag{1}$$

We adopt the same ground truth scaling to match S_H to S_L, using the inverse of the calculated scale from Equation 1 to get $S_{\mathrm{H}}^* = \frac{1}{c} S_{\mathrm{H}}$.

The preprocessing for the refinement training follows the same steps for the Hypersim [18] and the MultiIllum [16] datasets, except that we only require the HDR RGB as ground truth. For this reason, we can add additional datasets [10, 12, 13] to the training distribution which do not provide HDR intrinsic ground truth. For these datasets, we apply the full preprocessing pipeline from [13] to synthesize quantized, non-linear JPEGs and generate the inputs to our intrinsic reconstruction block by applying the fully trained dequantization and linearization networks from [13] on these JPEGs. More specifically, we add Gaussian noise to the raw images before the exposure, apply a randomly sampled CRF from [7] to the clipped RGB image, and write the results as 8-bit JPEG to disk.

We train our reconstruction networks for 1 mil. iterations each. After 500,000 iterations, we duplicate the weights and use the second, frozen set as a base to train the refinement network for 500,000 iterations in total, updating its reconstruction base every 100,000 iterations with the most recent weight set.

E Runtime Analysis

The formulation of our method in the intrinsic space requires a decomposition network as an additional step in the HDR reconstruction pipeline. To evaluate the impact of this modification in terms of computing efficiency, we compare the average run times per image in Table 8 against SingleHDR [13] which we use as our baseline. Since the processing steps up to the pixel reconstruction are identical for both cases, we only report the run times for the decomposition, reconstruction, and refinement networks, respectively.

We show the numbers for two datasets [9, 17] with different processing resolutions. For both cases, the decomposition pipeline including base and high-resolution estimation following Careaga *et al.* [2] is the most resource-intensive part of our approach. Our reconstruction and refinement networks, using the lightweight EfficientNet [20] architecture, have a small footprint on the run time. When compared to the HDR hallucination network of SingleHDR [13], our pipeline reduces the run time significantly.

Note that this reduction in the run time comes despite our use of multiple networks in our pipeline. This comes from the use of a heavy U-Net architecture in SingleHDR [13] versus our use of a lightweight EfficientNet [20]. Despite our use of simpler architectures, as our qualitative and quantitave analyses show, we are able to increase the performance while cutting down the required run time. We are able to successfully utilize smaller architectures thanks to our physically motivated method design that makes the dynamic reconstruction and color recovery tasks easier to model for a network.

References

1. Azimi, M., et al.: Pu21: A novel perceptually uniform encoding for adapting existing quality metrics for hdr. In: 2021 Picture Coding Symposium (PCS). pp. 1–5. IEEE (2021)
2. Careaga, C., Aksoy, Y.: Intrinsic image decomposition via ordinal shading. ACM Trans. Graph. **43**(1) (2023)
3. Chen, X., Liu, Y., Zhang, Z., Qiao, Y., Dong, C.: Hdrunet: Single image hdr reconstruction with denoising and dequantization. In: Proc. CVPR (2021)
4. Dang-Nguyen, D.T., Pasquini, C., Conotter, V., Boato, G.: Raise: a raw images dataset for digital image forensics. In: Proc. MMSys (2015)
5. Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R.K., Unger, J.: HDR image reconstruction from a single exposure using deep CNNs. ACM Trans. Graph. **36**(6) (2017)
6. Endo, Y., Kanamori, Y., Mitani, J.: Deep reverse tone mapping. ACM Trans. Graph. **36**(6) (2017)
7. Grossberg, M.D., Nayar, S.K.: What is the space of camera response functions? In: Proc. CVPR (2003)
8. Guo, C., Xiuhua, J.: Lhdr: Hdr reconstruction for legacy content using a lightweight dnn. In: Proc. ACCV (2022)
9. Hanji, P., Mantiuk, R., Eilertsen, G., Hajisharif, S., Unger, J.: Comparison of single image HDR reconstruction methods — the caveats of quality assessment. In: ACM Trans. Graph. (2022)
10. Kim, D., Kim, J., Nam, S., Lee, D., Lee, Y., Kang, N., Lee, H.E., Yoo, B., Han, J.J., Kim, S.J.: Large scale multi-illuminant (lsmi) dataset for developing white balance algorithm under mixed illumination. In: Proc. CVPR (2021)
11. Le, P.H., Le, Q., Nguyen, R., Hua, B.S.: Single-image hdr reconstruction by multi-exposure generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (January 2023)
12. Li, Z., Lu, M., Zhang, X., Feng, X., Asif, M.S., Ma, Z.: Efficient visual computing with camera RAW snapshots. IEEE Trans. Pattern Anal. Mach. Intell. (2024)
13. Liu, Y.L., Lai, W.S., Chen, Y.S., Kao, Y.L., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Single-image hdr reconstruction by learning to reverse the camera pipeline. In: Proc. CVPR (2020)
14. Mantiuk, R.K., Hammou, D., Hanji, P.: Hdr-vdp-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content. arXiv preprint arXiv:2304.13625 (2023)
15. Marnerides, D., Bashford-Rogers, T., Hatchett, J., Debattista, K.: Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In: Comput. Graph. Forum. vol. 37 (2018)
16. Murmann, L., Gharbi, M., Aittala, M., Durand, F.: A multi-illumination dataset of indoor object appearance. In: Proc. ICCV (2019)
17. Nemoto, H., Korshunov, P., Hanhart, P., Ebrahimi, T.: Visual attention in ldr and hdr images (2015), `http://infoscience.epfl.ch/record/203873`
18. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proc. ICCV (2021)
19. Santos, M.S., Ren, T.I., Kalantari, N.K.: Single Image HDR Reconstruction Using a CNN with Masked Features and Perceptual Loss. ACM Trans. Graph. **39**(4) (2020)

20. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proc. ICML (2019)
21. Zhang, L., Shen, Y., Li, H.: Vsi: A visual saliency-induced index for perceptual image quality assessment. IEEE Trans. Image Process. **23**(10), 4270–4281 (2014)

Table 5: Quantitative results against state-of-the-art on the HDR-Real [13] test split. We indicate an overlap with the training distribution of the full system with orange.

method	PSNR	VSI	HDR-VDP3	PSNR-H
OURS	32.62	95.65	**7.57**	25.82
DrTMO [6]	31.96	95.30	**7.57**	24.96
HDR-CNN [5]	30.64	92.13	6.70	22.62
ExpandNet [15]	30.69	94.45	6.97	23.86
Single-HDR [13]	**32.98**	**95.93**	7.47	**26.25**
Mask-HDR [19]	31.31	93.75	7.10	23.41
HDRUNet [3]	28.98	88.66	5.71	21.12
Multi-Exp Gen. [11]	31.59	95.19	7.34	25.24
Lightweight [8]	29.88	93.72	7.03	22.94

Table 6: Ablation results for different components of our reconstruction pipeline.

method	PSNR	VSI	HDR-VDP3	PSNR-H
full pipeline	36.77	98.28	8.95	32.83
w\o $\mathcal{L}_{MSG}$	36.68	98.10	8.93	32.56
w\o inf. mod	36.61	98.20	8.91	32.67

Table 7: Ablation results for the different components of our refinement network.

method	PSNR	VSI	HDR-VDP3	PSNR-H
full pipeline	36.77	98.28	8.95	32.83
w\o $\mathcal{L}_{MSG}$	36.19	98.09	8.91	31.91
w\o D_L and A_L	36.37	98.13	8.92	32.17
w\o $\hat{J}_H$	35.93	97.91	8.85	31.49
w\o add. datasets	36.30	98.11	8.93	32.22

Table 8: Runtime analysis of our method. We report the average run times (seconds) of our method in comparison with SingleHDR [13] on the HDR-Eye dataset [17] in 512 x 512 and on the SiHDR benchmark [9] in 1888 x 1280.

Method	Resolution	Decomposition	Reconstruction	Refinement	Total
Ours	512 x 512	0.10	0.02	0.01	0.12
SingleHDR [13]	512 x 512	-	0.67	0.21	0.89
Ours	1888 x 1280	0.32	0.02	0.03	0.34
SingleHDR [13]	1888 x 1280	-	0.69	0.22	0.91