GTA-HDR: A Large-Scale Synthetic Dataset for HDR Image Reconstruction

Supplementary Material

Hrishav Bakul Barua
Monash University & TCS Research
hrishav.barua@monash.edu

Kalin Stefanov
Monash University
kalin.stefanov@monash.edu

KokSheik Wong
Monash University
wong.koksheik@monash.edu

Abhinav Dhall
Monash University
abhinav.dhall@monash.edu

Ganesh Krishnasamy
Monash University
ganesh.krishnasamy@monash.edu

1. Related Work

This section complements Section 2 in the main paper. It provides further details on inverse tone mapping and a concise review of non-learning approaches for inverse tone mapping. The section also offers an overview of the no-reference quality assessment research which could benefit from the proposed GTA-HDR dataset.

1.1. Inverse Tone Mapping

Tone mapping [1, 11, 25] is the process of mapping the colors of HDR images capturing real-world scenes with a wide range of illumination levels to LDR images appropriate for standard displays with limited dynamic range. Inverse tone mapping [28] is the reverse process accomplished with either traditional non-learning methods or data-driven learning-based approaches. Fig. 1 illustrates an overview of the tone mapping pipeline and the process of inverse tone mapping using a data-driven model. Here, E is the sensor irradiance and Δt is the exposure time. The function $f_{crf}(E\Delta t)$ represents the tone mapping process, which outputs I_{LDR} images given I_{HDR} images captured by the camera sensor. The main goal of any HDR image reconstruction technique is to reverse the tone mapping process using another function $f_{crf}^{-1}(I_{LDR})/\Delta t$, which outputs reconstructed $\hat{I_{HDR}}$ images given I_{LDR} images. The main challenge is that the steps in $f_{crf}(E\Delta t)$ are generally not reversible [17]. We can, however, approximate the reverse process with a data-driven model $f_{DL}(I_{LDR}, \Theta)$, which reconstructs $\hat{I_{HDR}}$ images given I_{LDR} images, where Θ denotes the model parameters.

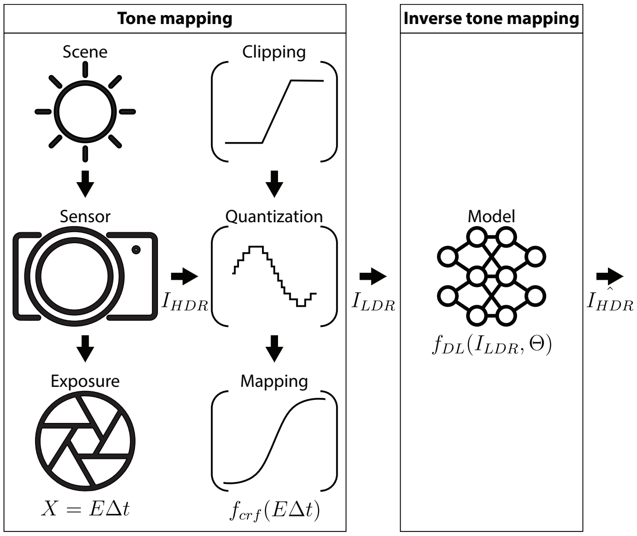


Figure 1. **Tone mapping and inverse tone mapping processes.** The camera function X is the product of the sensor irradiance E and exposure time Δt. The standard image formation pipeline (tone mapping) can be modeled with the function $f_{crf}(X)$, where $X = E\Delta t$. The goal of a data-driven inverse tone mapping model is to learn the function $f_{DL}(I_{LDR}, \Theta)$, where Θ are the model parameters, which correctly approximates the inverse of $f_{crf}(X)$.

1.2. Non-Learning Methods

Luzardo *et al.* [20] described an inverse tone mapping operator that allows higher peak brightness (*i.e.*, over 1000 nits) while converting LDR to HDR images. The process helps preserve the artistic intent of the reconstructed HDR images. Kovaleski and Oliveira [16] focused on enhancing the over/underexposed regions of images using cross-bilateral filtering. Huo *et al.* [12] presented an inverse tone mapping technique based on the human visual system. The approach uses human retina response to model the inverse local retina response using local luminance adaptation in the image. Masia *et al.* [21] addressed the ill-exposed areas of input LDR images, which are more prone to generate artifacts. This method uses an automatic global reverse tone mapping operator based on Gamma expansion along with automatic parameter calculation based on image statistics. Bist *et al*. [7] proposed a Gamma correction-based approach that adapts to the target lighting styles of the images. This work also added a color correction-based operator that reconstructs the intended colors in the HDR image.

1.3. No-Reference Quality Assessment

Due to the lack of ground truth HDR images and the high costs of resource-demanding full-reference quality metrics [2] such as High Dynamic Range Visual Differences Predictor [22], research has also accelerated in the field of no-reference HDR image quality assessment [5, 29]. Some approaches approximate the High Dynamic Range Visual Differences Predictor score for perceptual quality assessment of reconstructed HDR images using data-driven methods, such as CNN [3–5]. Other approaches approximate general quality score metrics of images (HDR or LDR), such as Peak Signal-to-Noise Ratio and Structural Similarity Index Measure, using CNN and distortion maps [26].

To address the data gap for no-reference image quality assessment, the GTA-HDR dataset contributes a set of distorted HDR along with the ground truth HDR and LDR images. The distorted HDR images can be utilized to develop no-reference quality assessment methods, *e.g*., by adopting a methodology similar to the ones proposed in [3–5]: 1) Estimate the full-reference quality scores for pairs of ground truth and distorted HDR images using an existing metric such as PSNR, SSIM, HDR-VDP-2/-3, and LPIPS; 2) Develop a data-driven method using the full-reference quality scores and their corresponding distorted HDR images; and 3) Utilize the developed model to estimate the quality scores of unseen reconstructed HDR images (*i.e*., no-reference quality assessment). Similarly, one can develop data-driven methods to estimate the quality scores for tone-mapped LDR and HDR images.

2. Results

This section complements Section 5 in the main paper. It provides qualitative results to further demonstrate the impact of GTA-HDR on the state-of-the-art in HDR reconstruction as well as in other computer vision tasks, including 3D human pose estimation, human body part segmentation, and holistic scene segmentation.

2.1. HDR Reconstruction

Fig. 2 illustrates examples of HDR images reconstructed by training ArtHDR-Net [6] with the GTA-HDR data in an end-to-end fashion. The histograms of the ground truth and the reconstructed images are also included. We can see that the histograms from the method trained with GTA-HDR data (*i.e*., HDR_{Ours}) are more similar to the histograms of ground truth HDR images (*i.e*., HDR_{GT}) than those from the method trained without GTA-HDR data (*i.e*., HDR_{Base}). We also report the Kullback-Leibler (KL) divergence values for tone-mapped HDR_{GT} and tone-mapped HDR_{Base} and HDR_{Ours} using the RGB intensities. We see the average KL divergence of the RGB histogram intensity distributions are significantly lower for HDR_{Ours} compared to HDR_{Base}.

Fig. 3 illustrates further qualitative results from the state-of-the-art method ArtHDR-Net [6] on extremely under/overexposed images. Similarly, Fig. 4 demonstrates the performance for arbitrary real images from the Internet. Both these cases show that GTA-HDR trained model is capable of recovering extremely over/underexposed images with great fidelity. To further illustrate the contribution of the GTA-HDR dataset on in-the-wild HDR image reconstruction, in Fig. 5 we show the results on two images selected from HDR-Real [19] dataset having extreme lighting, color, and contrast variations. We also report the PSNR, SSIM, and HDR-VDP-2 scores.

2.2. Downstream Applications

To further demonstrate the contribution of the proposed GTA-HDR dataset, this section illustrates its impact on the state-of-the-art in other computer vision tasks including 3D human pose estimation, human body part segmentation, and holistic scene segmentation.

2.2.1 3D Human Pose and Shape Estimation

We used BEV [27] as a state-of-the-art pre-trained 3D human pose and shape estimator from images. We tested the BEV model on the reconstructed HDR images from several versions of the state-of-the-art method ArtHDR-Net [6]. Table 1 reports the impact of the image pre-processing step (utilizing reconstructed HDR images from different versions of ArtHDR-Net) on BEV performance evaluated on the AGORA [23] 3D human pose dataset. We report two commonly used metrics, F1 Score to measure detection accuracy and Mean Per Joint Position Error (MPJPE) to measure pose accuracy. The results demonstrate that the pre-processing step enables a significant increase in the performance of BEV. In Fig. 6a we provide qualitative results in support of this quantitative evaluation.

2.2.2 2D Human Body Part Segmentation

In this experiment, we used CDCL [18], a state-of-the-art body part segmentation model. Similar to the previous case, we tested the model on reconstructed HDR images from several versions of ArtHDR-Net [6]. Table 1 reports the impact of the HDR reconstruction step on the COCO-DensePose [10] dataset, which is used for CDCL performance evaluation. We use the Mean Intersection of Union (mIOU%), *i.e*., the mean of all IoUs between predicted and ground truth masks to measure the accuracy of the predictions. We report the average accuracy for all the body parts

Table 1. **Impact of the GTA-HDR dataset on the performance of the state-of-the-art in 3D human pose and shape estimation, 2D human body part segmentation, and semantic segmentation.** We used ArtHDR-Net [6] trained with different datasets for HDR image reconstruction. The resulting HDR images from 1) AGORA [23] 3D human pose dataset were used by BEV [27] for 3D human pose and shape estimation; 2) COCO-DensePose [10] dataset were used by CDCL [18] for 2D human body part segmentation; and 3) Cityscapes [8] dataset were used by SAM [15] for semantic segmentation. *R* $\oplus$ *S*: Real and synthetic data combines all five datasets [9, 13, 14, 24, 30]; *GTA-HDR*: Proposed synthetic dataset; *None*: Results without HDR image reconstruction.

		3D human pose and shape estimation		**2D human body part segmentation**	**Semantic segmentation**
Pre-processing	**Datasets**	**F1 (detection)↑**	**MPJPE (pose)↓**	**mIOU%↑**	**mIOU%↑**
None	-	0.57	129	66.24	54.24
ArtHDR-Net	-	0.58	128.7	67.12	54.27
ArtHDR-Net	R $\oplus$ S	0.58	128.5	67.9	54.26
ArtHDR-Net	GTA-HDR	0.61	125.4	69.55	54.29
ArtHDR-Net	R $\oplus$ S $\oplus$ GTA-HDR	**0.65**	**121.9**	**74.71**	**54.36**

considered in [18]. The results establish the advantages of using the proposed pre-processing (*i.e.*, HDR reconstruction) step. Fig. 6b illustrates the contribution of the GTA-HDR dataset to this task. The human body part segmentation results are more accurate with the reconstructed HDR images than the over/underexposed LDR images. For the overexposed LDR images, one person is completely missed in the second image. For the underexposed LDR images, the output is noisy and erroneous.

2.2.3 Semantic Segmentation

Finally, we report an experiment on another vision application, *i.e.*, holistic semantic segmentation of scenes, which is an important task in robotics and human-robot interaction. We consider a recent state-of-the-art method SAM [15] as a pre-trained holistic scene segmentation model. Table 1 reports the improvements in the SAM output using the HDR reconstruction as a pre-processing step with the Cityscapes [8] dataset. We use Mean Intersection of Union (mIOU%) as the accuracy measure for segmentation. The results show steady improvements in the performance of SAM. Fig. 6c illustrates the impact of the GTA-HDR trained model ArtHDR-Net [6]. The objects/buildings in the background are not segmented well in the overexposed LDR images. Similarly, in the underexposed LDR images, even the near objects are sometimes segmented erroneously.

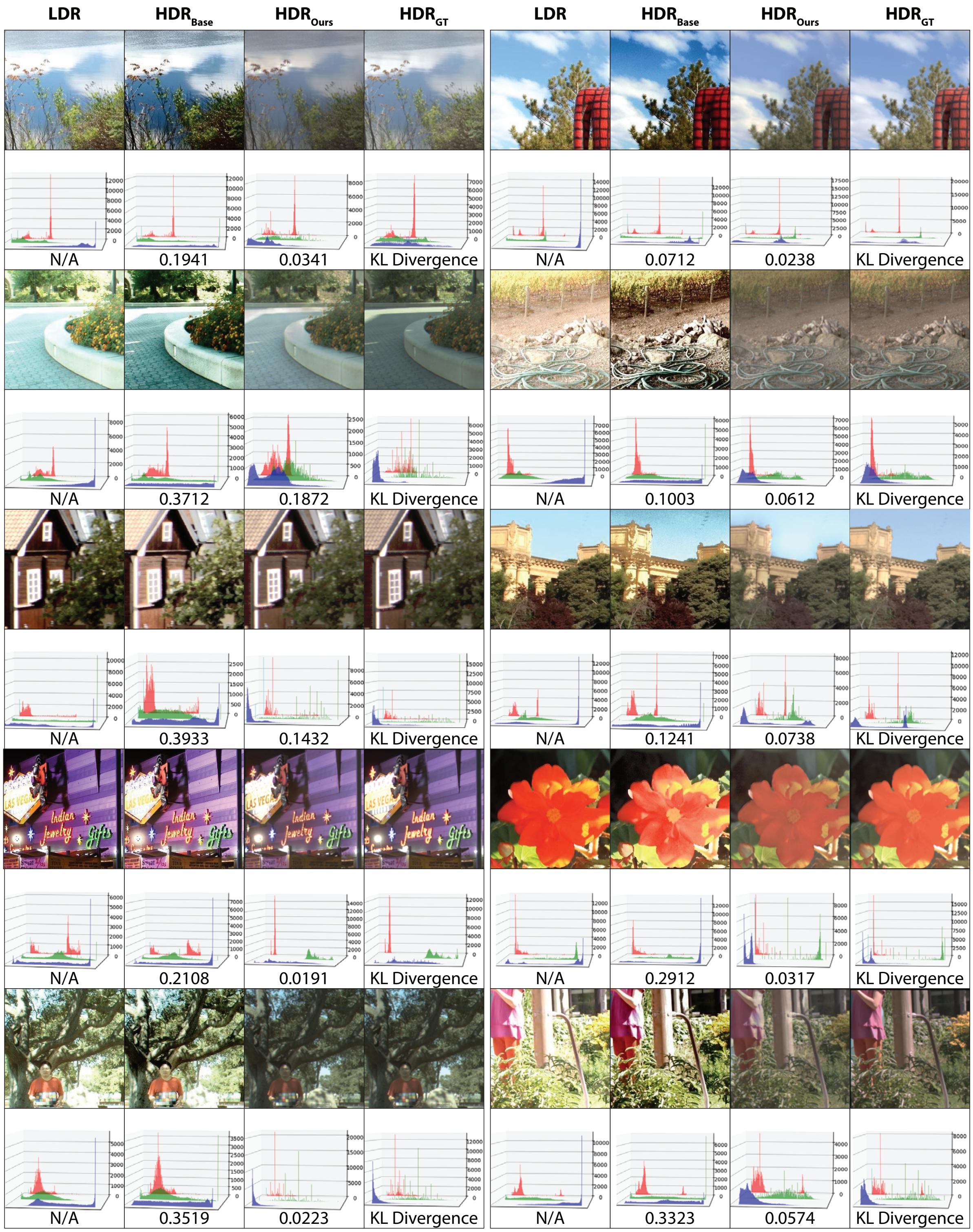


Figure 2. **HDR images reconstructed with and without GTA-HDR as part of the training dataset, along with the RGB histograms and KL-divergence values.** *Base*: HDR images reconstructed with ArtHDR-Net [6] trained without GTA-HDR data; *Ours*: HDR images reconstructed with ArtHDR-Net trained with GTA-HDR data; *GT*: Ground truth.

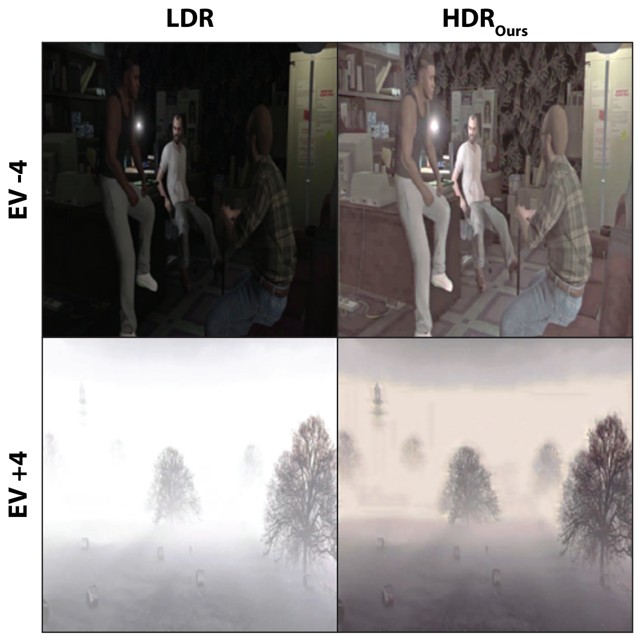


Figure 3. **Performance of ArtHDR-Net [6].** The state-of-the-art method was trained with the GTA-HDR dataset and used for HDR image reconstruction from highly overexposed and underexposed synthetic LDR images. *EV*: Exposure value; *Ours*: HDR images reconstructed with ArtHDR-Net trained with GTA-HDR data.

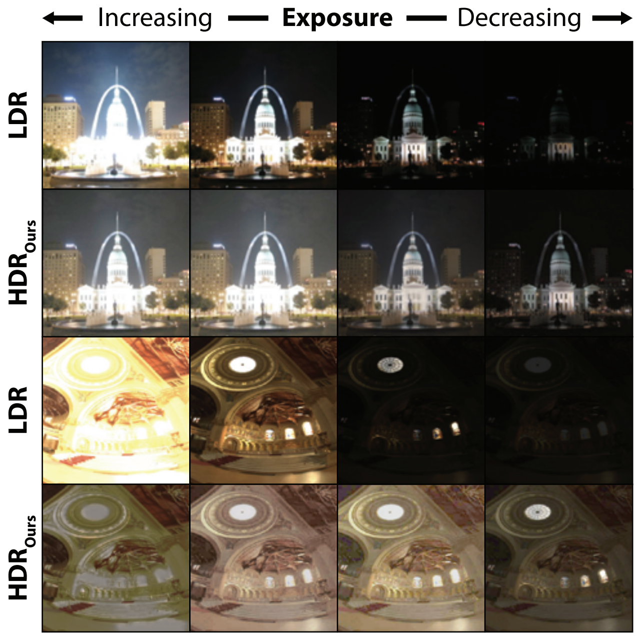


Figure 4. **Performance of ArtHDR-Net [6].** The state-of-the-art method was trained with the GTA-HDR dataset and used for HDR image reconstruction from highly overexposed and underexposed real LDR images from the Internet. *Ours*: HDR images reconstructed with ArtHDR-Net trained with GTA-HDR data.

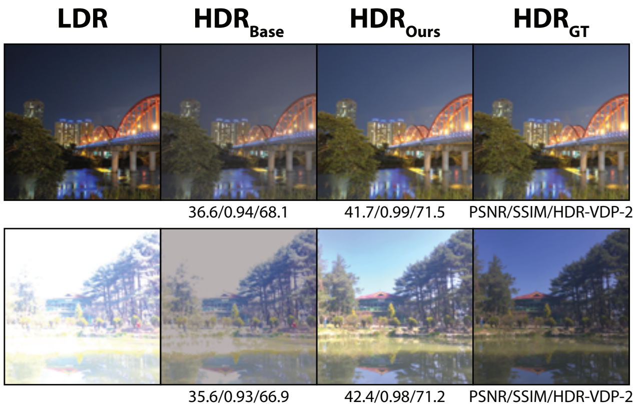
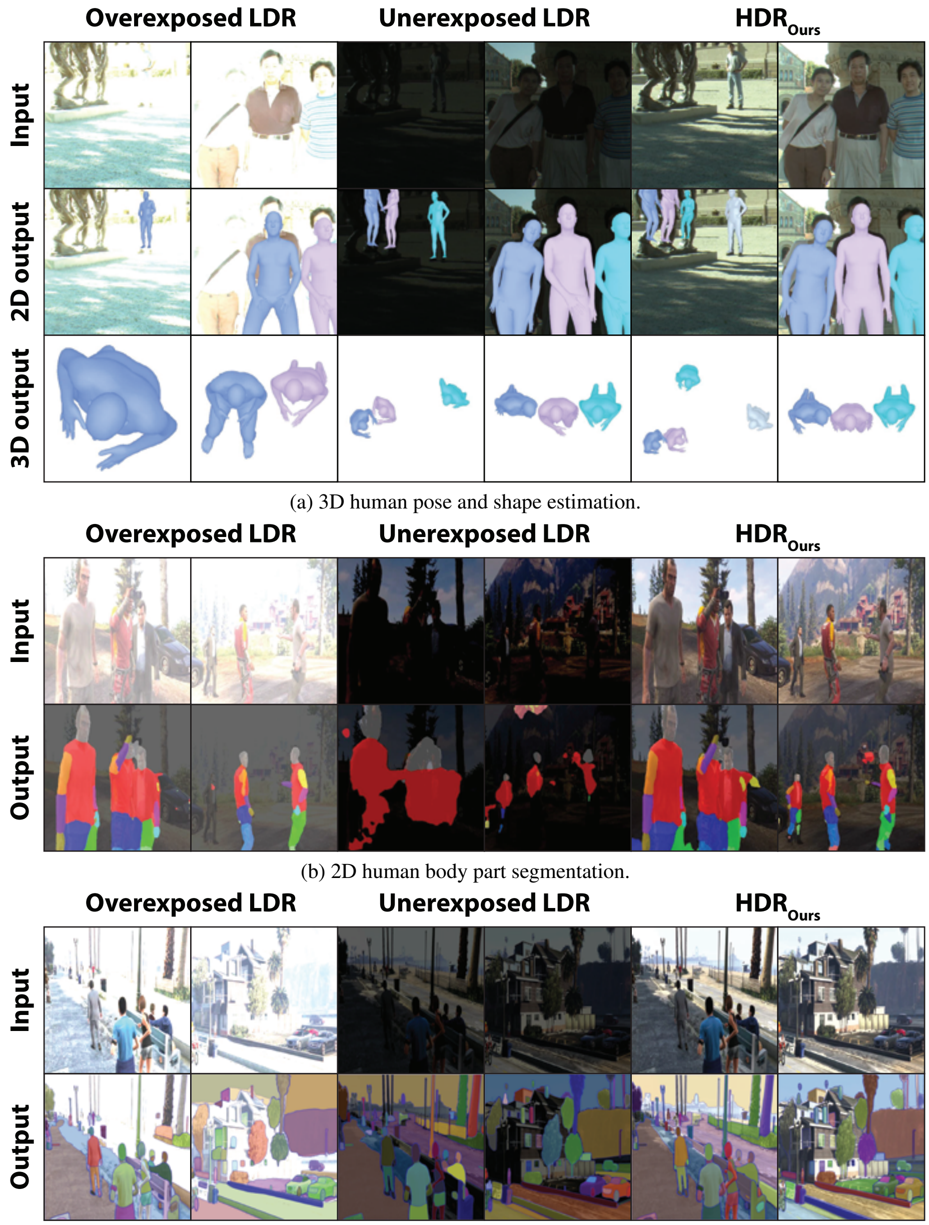


Figure 5. **Performance of ArtHDR-Net [6].** We show the results on two extreme real in-the-wild images selected from HDR-Real [19] dataset. These images have extreme lighting conditions, color variations, and contrast levels. *Base*: HDR images reconstructed with ArtHDR-Net trained without GTA-HDR data; *Ours*: HDR images reconstructed with ArtHDR-Net trained with GTA-HDR data; *GT*: Ground truth.



(a) 3D human pose and shape estimation.

(b) 2D human body part segmentation.

(c) Semantic segmentation.

Figure 6. **Impact of GTA-HDR on the performance of the state-of-the-art in 3D human pose and shape estimation, 2D human body part segmentation, and semantic segmentation.** We used ArtHDR-Net [6] trained with the GTA-HDR dataset for HDR image reconstruction. The resulting HDR images were used by BEV [27] (3D human pose and shape estimation), CDCL [18] (2D human body part segmentation), and SAM [15] (semantic segmentation). *Ours*: HDR images reconstructed with ArtHDR-Net trained with GTA-HDR.

References

[1] Ali Ak, Abhishek Goswami, Wolf Hauser, Patrick Le Callet, and Frédéric Dufaux. RV-TMO: Large-Scale Dataset for Subjective Quality Assessment of Tone Mapped Images. *IEEE Transactions on Multimedia*, 2022. 1

[2] Theyab Alotaibi, Ishtiaq R Khan, and Farid Bourennani. Quality Assessment of Tone-mapped Images Using Fundamental Color and Structural Features. *IEEE Transactions on Multimedia*, 2023. 2

[3] Alessandro Artusi, Francesco Banterle, Fabio Carra, and Alejandro Moreno. Efficient Evaluation of Image Quality via Deep-Learning Approximation of Perceptual Metrics. *IEEE Transactions on Image Processing*, 29:1843–1855, 2019. 2

[4] Francesco Banterle, Alessandro Artusi, Alejandro Moreo, and Fabio Carrara. Nor-Vdpnet: A No-Reference High Dynamic Range Quality Metric Trained On Hdr-Vdp 2. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 126–130. IEEE, 2020. 2

[5] Francesco Banterle, Alessandro Artusi, Alejandro Moreo, Fabio Carrara, and Paolo Cignoni. NoR-VDPNet++: Real-Time No-Reference Image Quality Metrics. *IEEE Access*, 11:34544–34553, 2023. 2

[6] Hrishav Bakul Barua, Ganesh Krishnasamy, KokSheik Wong, Kalin Stefanov, and Abhinav Dhall. ArtHDR-Net: Perceptually Realistic and Accurate HDR Content Creation. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 806–812. IEEE, 2023. 2, 3, 4, 5, 6

[7] Cambodge Bist, Rémi Cozot, Gérard Madec, and Xavier Ducloux. Tone expansion using lighting style aesthetics. *Computers & Graphics*, 62:77–86, 2017. 2

[8] 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*, pages 3213–3223, 2016. 3

[9] Lalonde et al. The Laval HDR sky database. http://hdrdb.com/, 2016. [Online; accessed 3-July-2023]. 3

[10] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 2, 3

[11] Xueyu Han, Ishtiaq Rasool Khan, and Susanto Rahardja. High Dynamic Range Image Tone Mapping: Literature review and performance benchmark. *Digital Signal Processing*, page 104015, 2023. 1

[12] Yongqing Huo, Fan Yang, Le Dong, and Vincent Brost. Physiological inverse tone mapping based on retina response. *The Visual Computer*, 30:507–517, 2014. 1

[13] Hanbyol Jang, Kihun Bang, Jinseong Jang, and Dosik Hwang. Dynamic Range Expansion Using Cumulative Histogram Learning for High Dynamic Range Image Generation. *IEEE Access*, 8:38554–38567, 2020. 3

[14] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep High Dynamic Range Imaging of Dynamic Scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017. 3

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3, 6

[16] Rafael P Kovaleski and Manuel M Oliveira. High-Quality Reverse Tone Mapping for a Wide Range of Exposures. In *2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 49–56. IEEE, 2014. 1

[17] Phuoc-Hieu Le, Quynh Le, Rang Nguyen, and Binh-Son Hua. Single-Image HDR Reconstruction by Multi-Exposure Generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4063–4072, 2023. 1

[18] Kevin Lin, Lijuan Wang, Kun Luo, Yinpeng Chen, Zicheng Liu, and Ming-Ting Sun. Cross-domain complementary learning using pose for multi-person part segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3):1066–1078, 2020. 2, 3, 6

[19] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-Image HDR Reconstruction by Learning to Reverse the Camera Pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1651–1660, 2020. 2, 5

[20] Gonzalo Luzardo, Jan Aelterman, Hiep Luong, Wilfried Philips, Daniel Ochoa, and Sven Rousseaux. Fully-Automatic Inverse Tone Mapping Preserving the Content Creator's Artistic Intentions. In *2018 Picture Coding Symposium (PCS)*, pages 199–203. IEEE, 2018. 1

[21] Belen Masia, Ana Serrano, and Diego Gutierrez. Dynamic range expansion based on image statistics. *Multimedia Tools and Applications*, 76:631–648, 2017. 1

[22] Manish Narwaria, Rafal K Mantiuk, Mattheiu Perreira Da Silva, and Patrick Le Callet. HDR-VDP-2.2: A calibrated method for objective quality prediction of high dynamic range and standard images. *Journal of Electronic Imaging*, 24(1):010501–010501, 2015. 2

[23] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 2, 3

[24] K Ram Prabhakar, Rajat Arora, Adhitya Swaminathan, Kunal Pratap Singh, and R Venkatesh Babu. A Fast, Scalable, and Reliable Deghosting Method for Extreme Exposure Fusion. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8. IEEE, 2019. 3

[25] Aakanksha Rana, Giuseppe Valenzise, and Frederic Dufaux. Learning-based tone mapping operator for efficient image matching. *IEEE Transactions on Multimedia*, 21(1):256–268, 2018. 1

[26] Chandra Sekhar Ravuri, Rajesh Sureddi, Sathya Veera Reddy Dendi, Shanmuganathan Raman, and Sumohana S Channappayya. Deep no-reference tone mapped image quality assessment. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 1906–1910. IEEE, 2019. 2

[27] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting People in their Place: Monocular Regression of 3D People in Depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022. 2, 3, 6

[28] Lin Wang and Kuk-Jin Yoon. Deep Learning for HDR Imaging: State-of-the-Art and Future Trends. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8874–8895, 2021. 1

[29] Bo Yan, Bahetiyaer Bare, and Weimin Tan. Naturalness-aware deep no-reference image quality assessment. *IEEE Transactions on Multimedia*, 21(10):2603–2615, 2019. 2

[30] Jinsong Zhang and Jean-François Lalonde. Learning High Dynamic Range from Outdoor Panoramas. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4519–4528, 2017. 3