Recasting Regional Lighting for Shadow Removal

Anonymous submission

In this supplemental, we first provide an in-depth explanation of the network structure and hyper-parameter settings. Subsequently, we expound on the mask labeling process employed for the SRD dataset. We then provide several discussions to address potential concerns. Finally, we conclude by presenting additional qualitative comparisons and results.

Configurations

In our Bilateral Correction Network, an encoder, consisting of two parallel branches for reflectance and corrected illumination layers, initially extracts cross-scale features. Each branch follows an architecture similar to the encoder in the shadow-aware decomposition network (as delineated in **Section 3.1** of the main submission), enhanced by two extra convolutional layers at each scale. Following feature extraction, we apply our Illumination-Guided Texture Restoration to these features to boost feature consistency by establishing a correlation between corrected illumination and reflectance features. Within IGTR, the local region size is set to 1/4 of the feature size in scales$\{1, 2, 3\}$ and 1/2 in scales$\{4, 5\}$. The shifting network detailed in Eq.(8) comprises two convolution layers, the first layer featuring a stride of $\{4, 4, 4, 2, 2\}$ at each of the five scales for spatial feature aggregation. The second layer, with an output dimension of 2 for vertical and horizontal offset, accomplishes channel aggregation. A subsequent scaling layer constrains the offset range within $4 \times [-1, 1]$, where 4 represents the four possible offset directions. The enriched features at each scale are progressively merged with the image features from the skip connections (as indicated by the gray dashed line in Fig.2 of the main submission), routed to the decoder, and ultimately result in the restoration of impaired shadow textures. The decoder architecture mirrors that utilized in the shadow-aware decomposition network.

Diffusion preliminaries.

Given a clean input image $\mathbf{x}_0$ and a timestep $t \in \{0, 1, ..., T\}$, the diffusion model (DDPM) (Ho, Jain, and Abbeel 2020) uses a forward diffusion Markov process to gradually add Gaussian noise to $\mathbf{x}_0$ to compute $\mathbf{x}_t$ satisfied $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I})$ at timestep t according to a predefined variance schedule β_t, where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. Then, $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}_t$ can be sampled via the parameterization trick, where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. During the reverse process, the diffusion model invert the forward process to generate a clean image $\mathbf{x}_0$ from random Gaussian noise satisfied $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$. With parameterization, μ_θ is obtained by training a noise prediction network $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ to predict the noise from $\mathbf{x}_t$, supervised by the ground truth sampled Gaussian noise $\boldsymbol{\epsilon}_t$: $\mathcal{L} = \mathbb{E}_{t\sim[1,T],\mathbf{x}_0,\boldsymbol{\epsilon}_t} \|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|_2^2$. For testing, the diffusion model sample random noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and iteratively denoises using $\boldsymbol{\epsilon}_\theta(\cdot)$ to get $\mathbf{x}_{t-1}$ from $\mathbf{x}_t$ until $t = 0$.

Inference flow of the LLC. Employing the decomposed shadow illumination $\mathbf{L}_\mathrm{s}$ as a conditional input, we commence at time t by sampling a random Gaussian noise map x_t. This is blended with the conditional input to form the time-embedded non-shadow lighting condition, $\mathbf{C}_t$. Subsequently, we concatenate $\mathbf{C}_t$ with $\mathbf{L}_\mathrm{s}$ along the channel axis, and this amalgamation (time embedding t is the default and required input, and is ignored here for simplification) serves as input to the LLC (*i.e.*, denoising network). After a series of denoising steps (*i.e.*, 50 steps), we acquire the corrected illumination, denoted as $\hat{\mathbf{L}}_\mathrm{s}$. Given that we implement the local conditional denoising process, we can initiate from a considerably small time step T with a comparatively larger end variance scheduler β_t (*e.g.*, 0.5). This allows for quicker inference compared to the original 1000-step denoising process. We adopt the improved UNet (Dhariwal and Nichol 2021) as our noise prediction and modify the input channel to 6 (*i.e.*, 3 for $\mathbf{L}_\mathrm{s}$ and 3 for $\mathbf{C}_\mathrm{t}$). Please refer to the original repository [1] for more network details.

Mask labeling on SRD dataset

This dataset is characterized by an assortment of complex shadow forms and scenes, encompassing tree branches, trunks, brooms, and panoramic views, coupled with a substantial number of soft shadows, thereby adding to the intricacies of labeling. When dealing with images featuring black-colored objects or soft-shadow-laden scenes, we utilize a contour-based methodology. This approach involves an initial outline of the penumbra's outer boundary, followed by a consistent inward transition aimed at minimizing the chance of false boundary labeling. In this way, annotation

[1] https://github.com/openai/guided-diffusion

Figure 1: Visual illustration of the SRD samples and our annotated masks. In each scene, we randomly selected two samples so that the user could clearly distinguish the shadow areas. Note that the mask area in the last sample is visually a bit too large in the upper left corner, but this is **reasonable** because that area is a soft shadow area and our labeled mask covers it completely.

(a) $\mathbf{I}_s$ (b) $\mathbf{R}_s$ (c) $\mathbf{L}_s$ (d) $\hat{\mathbf{L}}_s$ (e) $\hat{\mathbf{I}}_s$

Figure 2: Visual illustration of the impact of the shadow-aware decomposition on the final results. To circumvent the impact of color inconsistency between the input shadows and the ground truth during training, we employ the model weights trained on the ISTD+ dataset, which has partially alleviated this color-related issue. Please zoom in to see the shadow residual in (b).

masks can cover both hard and soft shadows. See Fig. 1 for visual illustrations. During the annotation phase, each sample is concurrently annotated by two professional annotators, with the degree of overlap between their annotations calculated. Should a sample's overlap rate exceed 90%, the annotations are averaged to yield the final output. In contrast, samples failing to meet this benchmark are submitted to additional annotation rounds. This process is iteratively performed until all samples satisfy the defined overlap rate. On average, the annotation process requires around 3 minutes per image, amounting to a total of 20 days for two professionals to fully annotate 3,088 shadow images.

Discussion

Exact shadow removal challenge. Shadow removal entails illumination correction and texture restoration within shadow regions. Conventional methods of learning color mappings fall short in adequately restoring degraded textures. Our pivotal discovery lies in the fact that texture restoration must be contingent upon the corrected local illumination. Our approach adeptly navigates the complexities of shadow-aware decomposition (where traditional decomposition methods may falter), local illumination correction, and conditional texture restoration.

Poor reflectance of $\mathcal{L}_{\text{fid}} + \mathcal{L}_{\text{ill}}$. Direct supervision of L with shadow annotations is unfeasible as we lack Ground Truths (GTs) for R and L. As such, we turn to the Retinex theory (Land 1977), assuming consistent reflectance across shadow and non-shadow images. We employ the first term of Eq.2 to diminish the disparities between $\mathbf{R}_s$ and $\mathbf{R}_{sf}$, while the second term serves to indirectly supervise $\mathbf{L}_s$ and $\mathbf{L}_{sf}$. Note that the first term of Eq.2 only constrains $\mathbf{R}_s$ and $\mathbf{R}_{sf}$ to be identical, yet ignore the details preservation and color correction. This omission explains the subpar visual quality of the reflectance result in Fig.8(c) of the main submission. Thus, we turn to Eq.3 to address this problem.

Effect of the Shadow-aware decomposition on final results. Based on the Retinex Theory, all shadow degradation should be confined to the illumination layer/image. However, the decomposition process can't be entirely reliable, as it lacks Ground Truth (GT) signals for training, despite the introduction of various physical regularizations to mitigate this issue. There might be concerns that decomposition failure could negatively impact the final result. As demonstrated

in Fig. 2, even when decomposition encounters partial failure (i.e., shadow residuals persist on the reflectance layer), our approach can proficiently manage the situation and produce correct results. This capability largely stems from our LLC's ability to generate correct illumination and the maintained efficiency of our texture restoration, which models the correct correspondence between the re-casted illumination and the decomposed reflectance.

Naive UNet is less effective for Local Lighting Correction. As highlighted in EMNet (Zhu et al. 2022), utilizing a regression-based network to learn a set of fixed parameters proves to be inadequate, given that shadows exhibit diverse shapes and non-uniform illumination. Instead, we employ the conditional diffusion model to iteratively generate accurate lighting for the local shadow region. The UNet-based method's efficacy is inferior to ours due to its limited generation capability. We further substantiate this viewpoint by performing local lighting correction via the naive UNet (Ronneberger, Fischer, and Brox 2015). This version only achieves an RMSE of 8.88 and PSNR of 33.36 within shadow regions on the ISTD dataset, lagging behind our method by 2.34 and 3.25 (see the Tab.6 of the main submission), respectively.

Experiments with different coefficients of the perceptual loss. We have also explored various combinations of loss ratios between pixel-wise and perceptual loss. As depicted in Table 1, while some other combinations yield slightly better values than ours in specific areas, for instance, the PSNR of (1 vs 1) exceeds ours by 0.22 in shadow regions, our finalized version (*i.e.*, 1 vs 0.1) still achieves superior performance across multiple metrics and regions. Note that we did not engage in careful adjustments of the loss ratios. It is plausible that better performance could be attained by experimenting with other combinations, or by utilizing Neural Architecture Search (NAS) to identify the optimal combination.

Ratios	RMSE ↓			PSNR ↑			SSIM ↑		
	S	*NS*	*All*	*S*	*NS*	*All*	*S*	*NS*	*All*
1*vs* 0.01	6.60	3.36	3.89	36.59	35.65	32.35	0.987	0.978	0.960
1*vs* 0.1(Ours)	**6.54**	**3.40**	**3.91**	**36.61**	**35.75**	**32.42**	**0.988**	**0.979**	**0.961**
1*vs* 1	6.58	3.77	4.23	36.83	34.92	32.14	0.988	0.975	0.958
1*vs* 10	6.53	3.60	4.08	36.68	35.43	32.37	0.987	0.977	0.959

Table 1: Quantitative comparisons with state-of-the-art shadow removal methods on the ISTD dataset.

Qualitative Comparisons

Comparisons to SOTAs.

Comparison on SRD. To ensure a fair and accurate comparison, we re-train existing methods using our manually annotated shadow masks, and report their performance and ours. A group of visual results are shown in the first row of Fig. 3, from which we can see that while existing methods generally recover the global illuminations partially, they are not able to correct the textures well. In contrast, our method can produce shadow-free image with correct textures.

Comparison on ISTD/ISTD+. We evaluate our proposed method on the ISTD and ISTD+ datasets using the shadow masks officially released by (Wang, Li, and Yang 2018), and display several visual comparison in Fig. 4. The 5th row shows a challenging case where the background surface contains dense chinese characters. Existing shadow removal methods cannot remove sharp shadow boundaries and produce inconsistent colors after removal. In contrast, our result has more consistent colors and does not have obvious shadow ghosting near the shadow boundaries due to our separate modeling of lighting and textures in shadow regions.

Comparison on LRSS. In Fig. 5, we present qualitative comparisons between our method and three state-of-the-art shadow removal methods on the LRSS (Gryka, Terry, and Brostow 2015) soft shadow dataset to further demonstrate the effectiveness of our approach. By default, all methods use the trained model on the SRD dataset for testing. Clearly, all competing methods fail to effectively remove the subtle soft shadows. Our method, on the other hand, produces quite realistic shadow-free results that accurately capture the true appearance of the scene.

Results on real-world samples.

Beyond the currently available shadow removal datasets, we display more visual predictions across diverse shadow scenes. Fig. 6 shows our method's predictions on several natural scenes exhibiting varied local illumination intensities.

More internal comparisons and analysis.

In Fig. 7, we display more decomposition comparisons to related shadow removal works and retinex-based methods. EMNet (Zhu et al. 2022) and SP+M+I-Net (Le and Samaras 2021) are two shadow removal works that utilize shadow-aware illumination maps. Both approaches directly consider the resulting shadow-free predictions as the reflectance layer and handle the lighting and object textures together, lacking constraints on the illumination map and resulting in shadow artifacts in their results. RetinexNet (Wei et al. 2018) and DeepUPE (Wang et al. 2019) are two low-light enhancement approaches that rely on the retinex model. However, they struggle to handle spatially inconsistent shadow scenes. Instead, our method shows the capability to effectively handle both homogeneous colors (first two rows) and complex texture details (last two rows), owing to the inclusion of shadow-aware regularization designs.

In Fig. 8, we display two more ablated examples to validate the designed loss function for shadow-aware decomposition. We emphasize, as stated in the main submission's Eq.(3), that $\mathbf{L}_{\text{s}}$ is not used. We exclusively employ $\mathbf{L}_{\text{sf}}$ given its absence of shadows. Utilization of $\mathbf{L}_{\text{s}}$ may misguide the decomposition model to concentrate excessively on the shadow regions, thereby leading to an undesired all-white coloration for non-shadow regions within the illumination layer. We've conducted a corresponding experiment to substantiate this claim. By augmenting $\mathcal{L}_{\text{ref}}$ with the $\mathbf{L}_{\text{s}}$ item, we present the results as $\mathcal{L}_{\text{fid}} + \mathcal{L}_{\text{ref}}$ (with $\mathbf{L}_{\text{s}}$). The resulting decomposed illumination layers, akin to the inverted shadow

matte [2] in De-shadowNet (Qu et al. 2017), display values nearing 255 in non-shadow regions. This consequently causes discernible shadow artifacts within the reflectance layer.

Fig. 9 exhibits additional shadow-aware decomposition results from our method, encompassing decomposed reflectance and illumination layers, reconstructed illumination layer, and our concluding predictions. All samples originate from the SBU (Vicente et al. 2016) shadow detection dataset. This not only showcases the adaptability of our method's shadow-aware decomposition outside the existing shadow removal datasets, but also the ultimate prediction capacity of our model.

References

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. In *NeurIPS*.

Gryka, M.; Terry, M.; and Brostow, G. J. 2015. Learning to remove soft shadows. *ACM TOG*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Land, E. H. 1977. The retinex theory of color vision. *Scientific American*.

Le, H.; and Samaras, D. 2021. Physics-based shadow image decomposition for shadow removal. *IEEE TPAMI*.

Qu, L.; Tian, J.; He, S.; Tang, Y.; and Lau, R. W. 2017. Deshadownet: A multi-context embedding deep network for shadow removal. In *CVPR*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Vicente, T. F. Y.; Hou, L.; Yu, C.-P.; Hoai, M.; and Samaras, D. 2016. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *ECCV*.

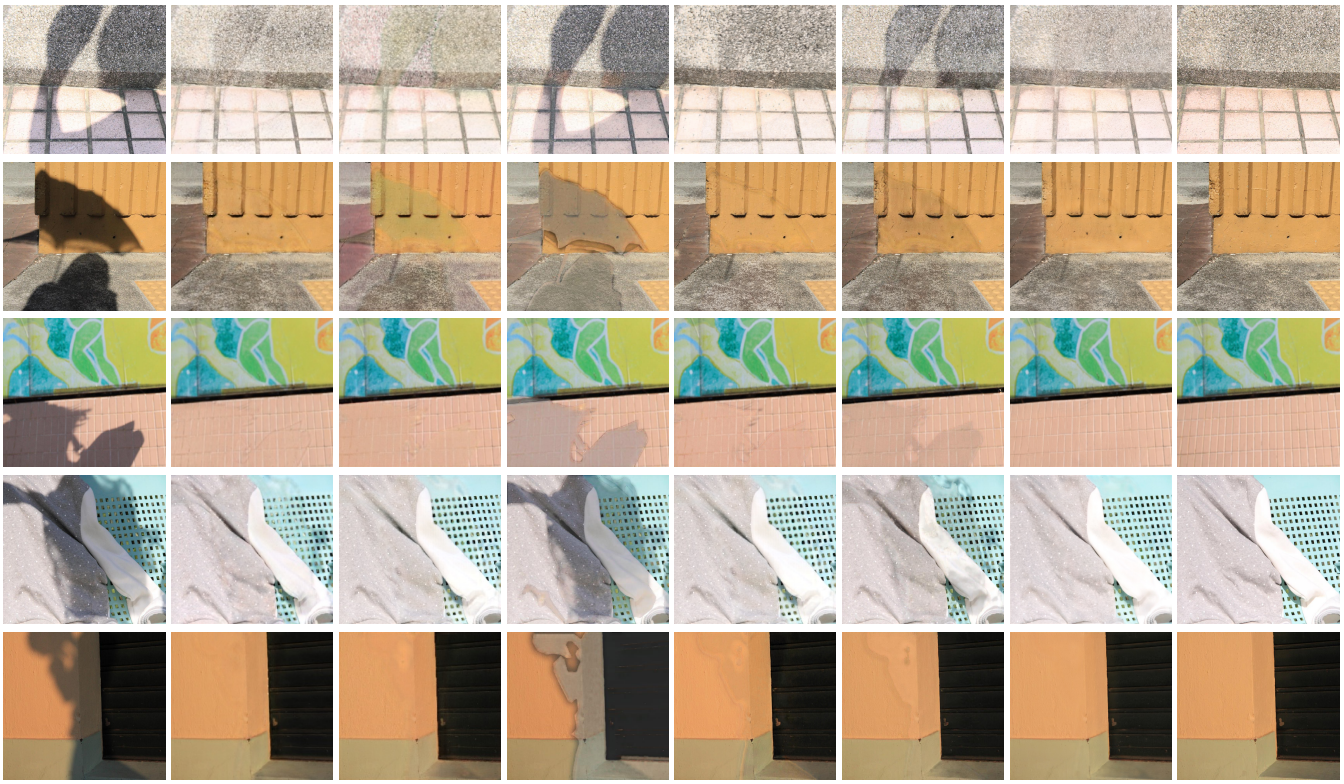
Wang, J.; Li, X.; and Yang, J. 2018. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*.

Wang, R.; Zhang, Q.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; and Jia, J. 2019. Underexposed photo enhancement using deep illumination estimation. In *CVPR*.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep retinex decomposition for low-light enhancement. In *BMVC*.

Zhu, Y.; Xiao, Z.; Fang, Y.; Fu, X.; Xiong, Z.; and Zha, Z.-J. 2022. Efficient Model-Driven Network for Shadow Removal. In *AAAI*.

[2] The original shadow matte formulation is: $\mathbf{I}_{\mathrm{s}} = \mathbf{S}_{\mathrm{m}} * \mathbf{I}_{\mathrm{sf}}$, where $\mathbf{S}_{\mathrm{m}}$ represents the shadow matte.



Input Shadow DC-GAN AEF G2R BMNet SGNet Ours GT

Figure 3: Qualitative comparisons with the state-of-the-art methods on the SRD dataset.

Input Shadow DC-GAN AEF G2R BMNet SGNet Ours GT

Figure 4: Qualitative comparisons with the state-of-the-art methods on the ISTD+ dataset.

Figure 5: Visual comparisons with state-of-the-art shadow removal methods on the LRSS datasets. Note that LRSS (Gryka, Terry, and Brostow 2015) doesn't provide corresponding shadow-free images and all results presented here employ weights trained on the SRD dataset. The image resolution is 1088×720.

Figure 6: Presented are the results of our method (model weights trained on the SRD dataset) for real-world shadow images not included in the existing shadow removal datasets. Images and results are arranged in pairs on two separate rows. The shadow examples were sourced from the SBU (Vicente et al. 2016) shadow detection dataset.

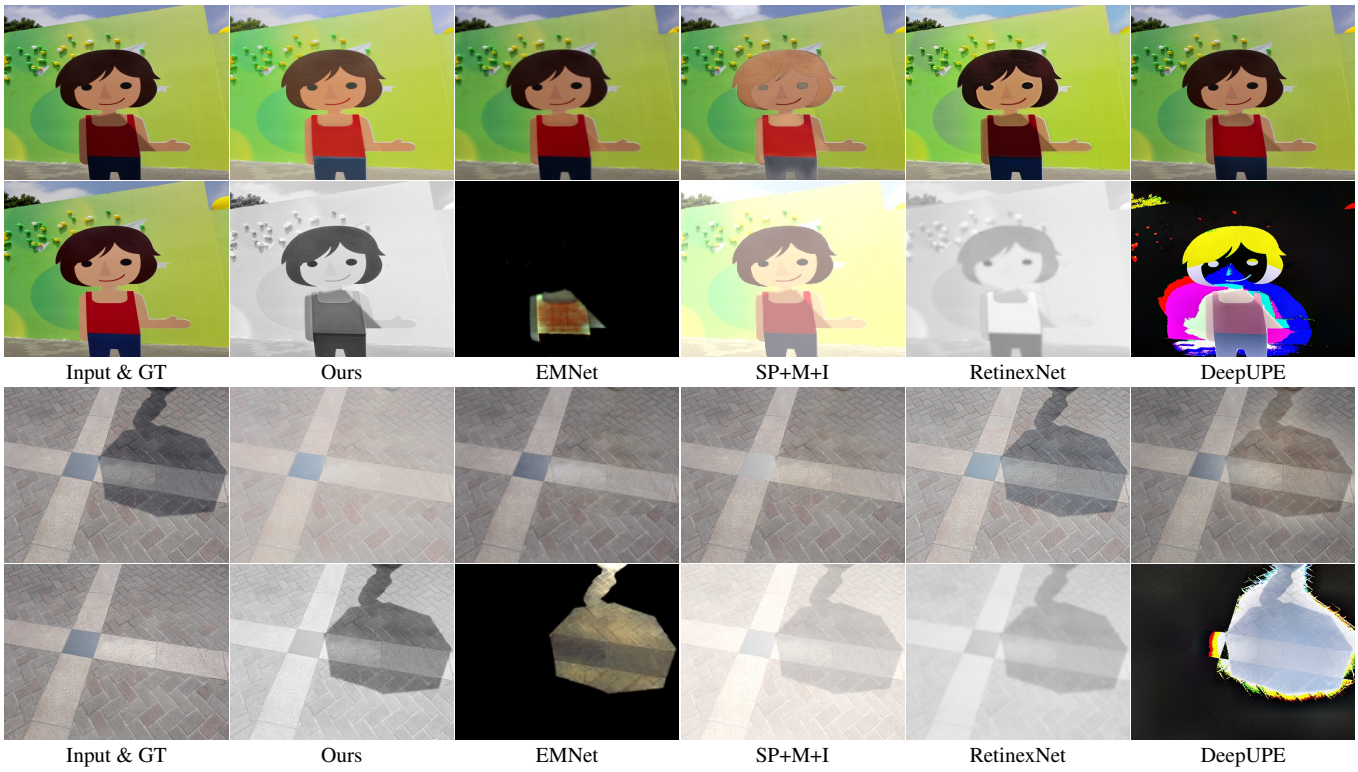


Figure 7: Visual comparisons of decomposition results among our method, two shadow formation model-based methods EMNet (Zhu et al. 2022) and SP+M+I (Le and Samaras 2021), and two retinex-based low-light enhancement methods RetinexNet (Wang et al. 2019) and DeepUPE (Wei et al. 2018). The output of EMNet and SP+M+I are their output shadow-free predictions (1st row) and intermediate shadow-aware illumination maps (2nd row).

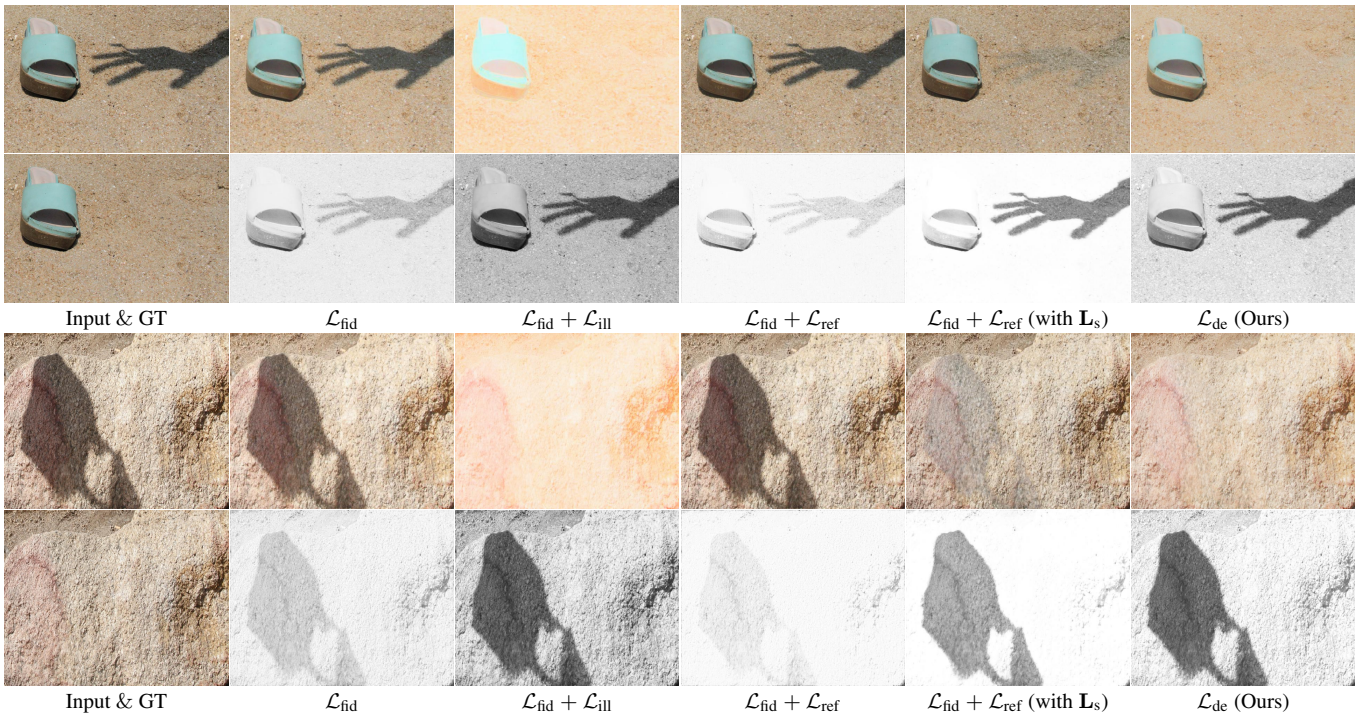


Figure 8: Visual comparisons of shadow decomposition of our method employing different regularizations.

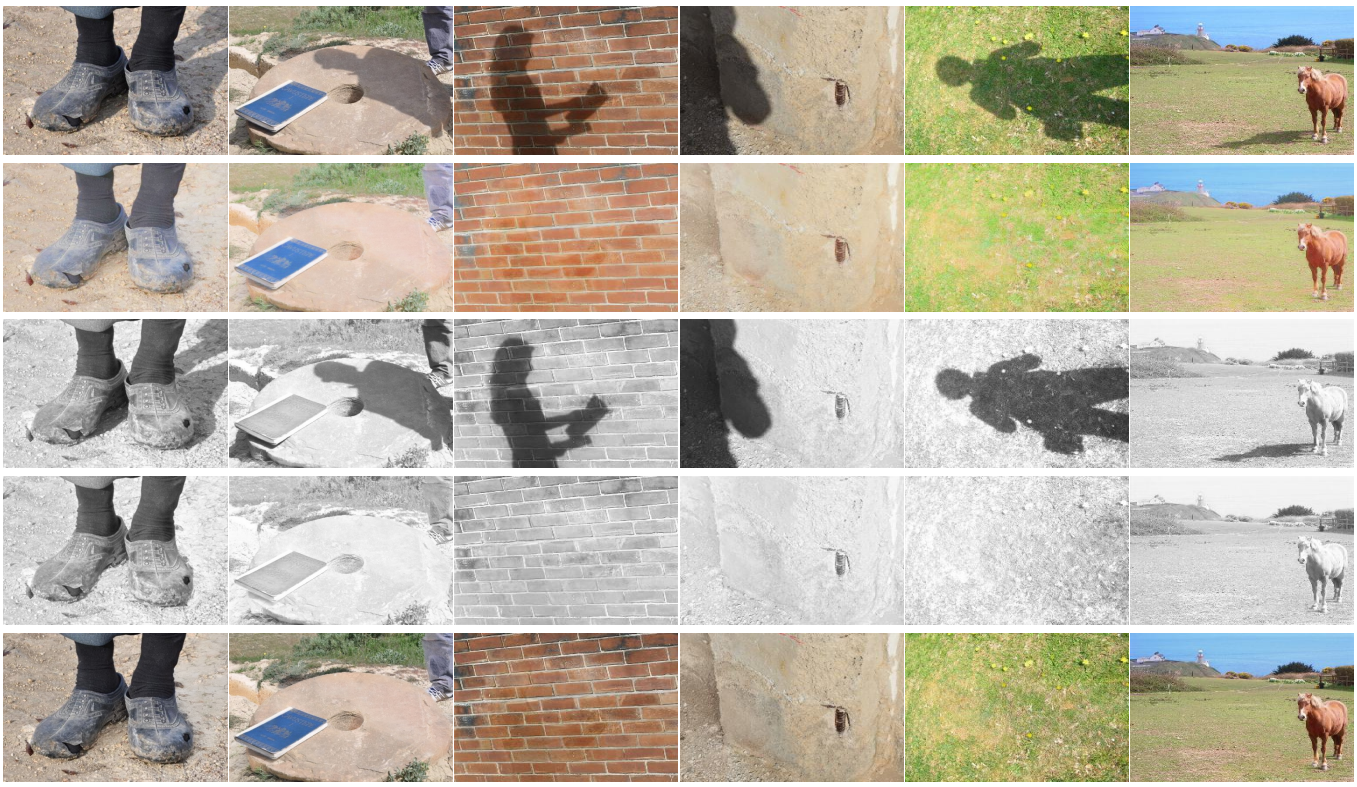


Figure 9: Displays of various outputs of our method. From top to bottom are input images (1st row), decomposed reflectance (2nd row) and illumination (3rd row) layers, our re-casted illumination (4th row) layers, and our shadow removal results (5th row). These shadow examples are sourced from the SBU (Vicente et al. 2016) shadow detection dataset, which lacks corresponding shadow-free images.