UnReflectAnything: RGB-Only Highlight Removal by Rendering Synthetic Specular Supervision

Supplementary Material

1. Extended Qualitative Inspection

We provide additional qualitative results of *UnReflectAnything* beyond those shown in the main paper. These examples further illustrate the model's behavior across diverse scenes and highlight its ability to suppress and inpaint specularities under varying appearance conditions. Fig. S2 shows additional input-output pairs, while in Fig. S3 we display an extended qualitative comparison across all relevant related work for several representative images from the general and surgical domains.

2. Challenging Scenarios

As discussed in the main paper, although *UnReflectAnything* generalizes well across multiple domains, some failure cases persist. We present illustrative examples in Fig. S1. A typical issue arises from the luminance-based detection of dataset highlights: highly reflective or bright surfaces, such as white anatomical structures in endoscopy, may be misinterpreted as specularities and subsequently inpainted (Fig. S1a). Still, a hard binary threshold for inpainting targets ensures that patches requiring inpainting are reliably identified, while the local mean prior $\mathbf{F}_{\text{mean}}$ preserves visible diffuse texture beneath highlights. We empirically tested soft masking, where downsampled I_{high} directly provides patch weights to indicate how much inpainting each patch requires (Tab. S1). We qualitatively assessed that the model only marginally modifies soft highlight regions, failing to remove highlights effectively. A similar phenomenon occurs in outdoor scenes (Fig. S1b), where very bright sky regions are incorrectly classified as highlights.

3. Architecture Design Choices

We conduct additional ablations to further justify the architectural and supervisory choices in *UnReflectAnything* (Table S1). Replacing MoGe-2 [2] normals with naïve depth-gradient normals markedly increases MSE_m, confirming that inaccurate normals produce unrealistic synthetic highlights and weaken supervision. Disabling token-space inpainting and relying solely on RGB-space reconstruction produces blurrier outputs and reduces *SSIM*, showing that restoring corrupted features before decoding is essential for preserving structure. Removing the local-mean prior further destabilizes token completion, particularly for large highlight regions. Similarly, positional encoding and learnable representations for inpainting targets [1] seem to benefit the model performance. Jointly training the decoder from scratch results in inferior performance compared to our two-stage curriculum, indicating that decoder pre-training offers a more stable feature-to-RGB initialization. Excluding dataset highlights from supervision is likewise crucial: supervising clipped pixels biases the model toward interpreting saturated reflections as diffuse regions, whereas masking them preserves a consistent and physically grounded learning signal. Table S2 reports the numeric values of all loss weights used during training, which remain fixed throughout optimization.

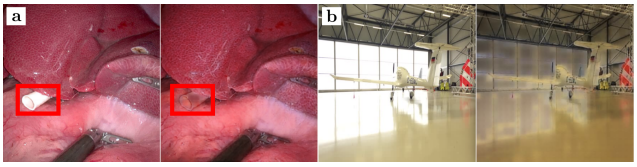

Figure S1. Representative failure modes of *UnReflectAnything*.

Table S1. **Ablation Studies:** Metrics are averaged across all datasets.

Configuration / Ablation Experiment	$MSE_m \downarrow$	*SSIM* ↑
Masking Strategy		
w/ Soft Mask conditioning	0.005	0.937
Supervision		
w/o MoGe-2 (depth-gradient normals)	0.012	0.909
Model Architecture		
w/o $\mathbf{F}_{\texttt{mean}}$ (Eq. 6, $\lambda = 1$)	0.004	0.943
w/o token inpainting (RGB inpainting)	0.007	0.816
w/o Positional Encoding	0.005	0.934
w/o $\mathbf{f}_{\texttt{mask}}$ (Eq. 6, $\lambda = 0$)	0.006	0.929
Training Curriculum		
w/o decoder pre-training	0.006	0.873
w/o dataset-highlight exclusion	0.022	0.933
Full Model (OURS)	**0.003**	**0.957**

Table S2. Loss function weights used at training time.

Loss Term	**Symbol**	**Value**
Highlight Dice loss	w_{dice}	0.2
Highlight L1 loss	w_{L1}	0.7
Highlight TV regularizer	w_{TV}	0.1
Seam loss	w_{seam}	0.25
Specularity penalty	w_{spec}	0.25
Diffuse RGB reconstruction	w_{RGB}	0.5

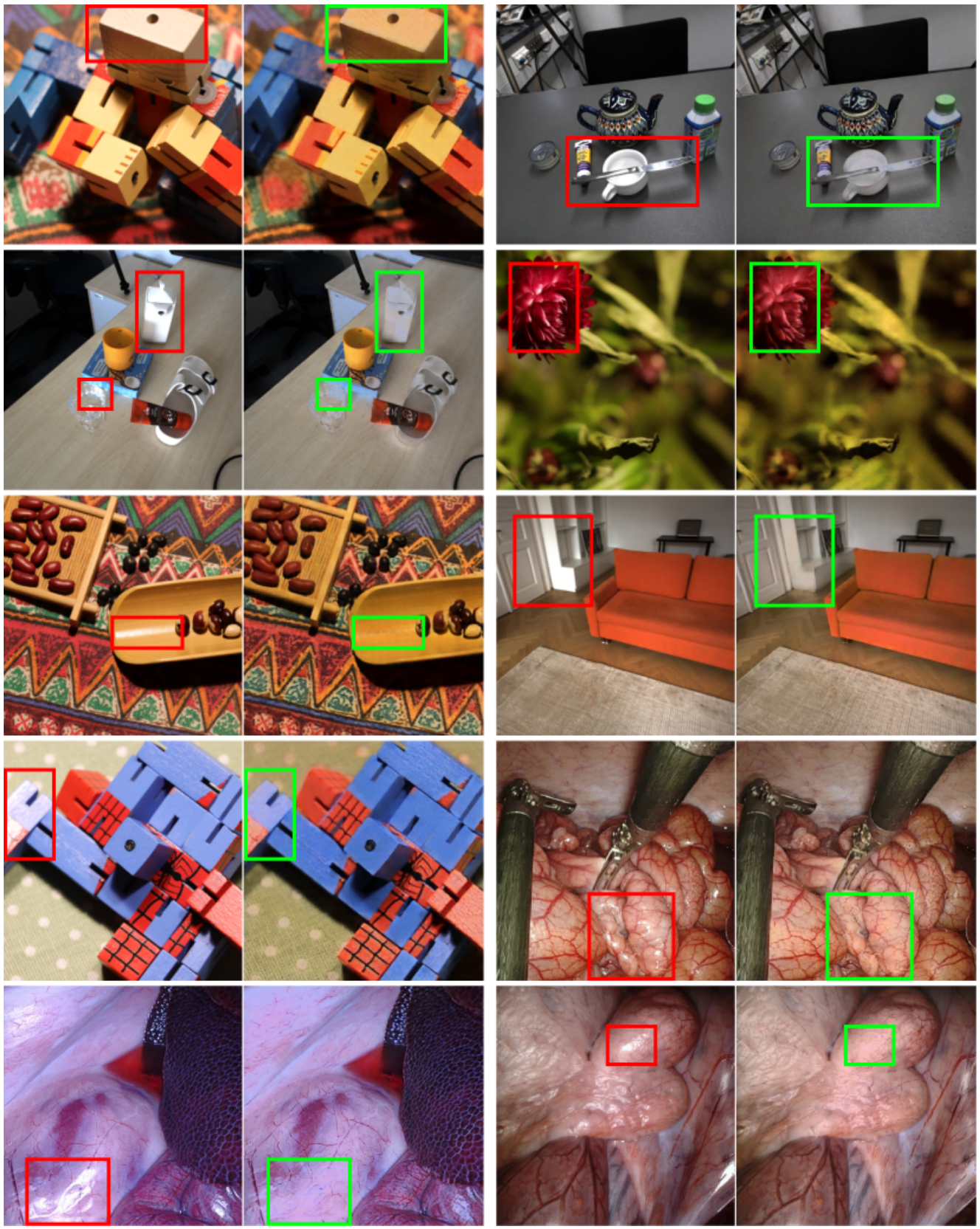

Figure S2. Input–output examples for *UnReflectAnything* across multiple datasets. We indicate highlights in the input images (left of each pair) with a red rectangle and the highlight-free reconstruction in the output image (right of each pair) with a green rectangle

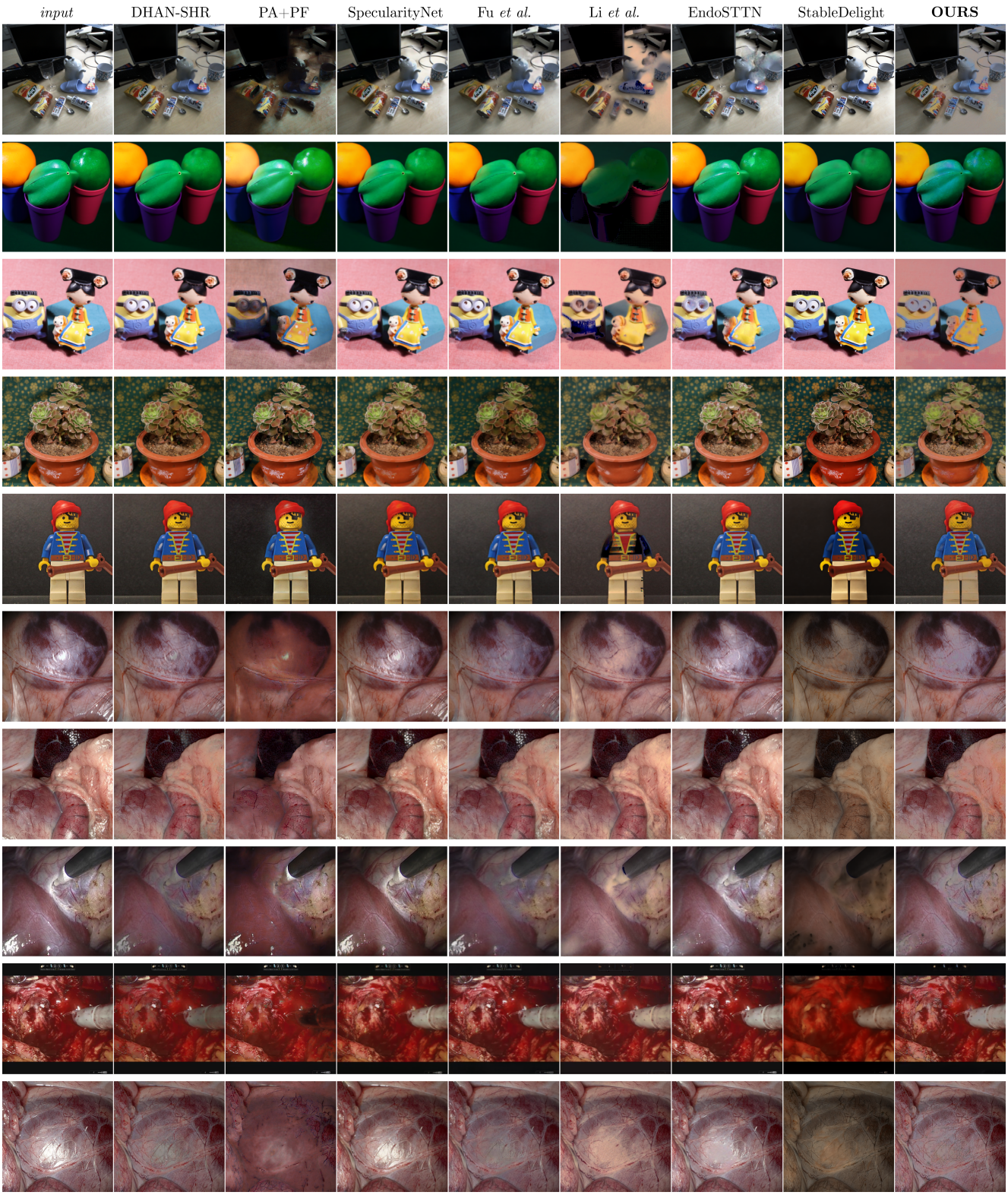

Figure S3. Extensive qualitative comparison of reflection-removed outputs on several images from all related work. Best viewed digitally.

References

[1] Yuxin Fang, Shusheng Yang, Shijie Wang, Yixiao Ge, Ying Shan, and Xinggang Wang. Unleashing vanilla vision transformer with masked image modeling for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6244–6253, 2023.

[2] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025.