PhaSR: Generalized Image Shadow Removal with Physically Aligned Priors

Supplementary Material

Overview

This supplementary material provides comprehensive details to support the main paper. The document is organized as follows:

- **Section 1: Data Loading and Preprocessing** – Details on depth-to-normal conversion, normal map normalization, and input preparation pipeline using DepthAnything-V2 and DINO-V2.
- **Section 2: Algorithm Description** – Complete algorithmic specification of the PhaSR training pipeline, including physically aligned normalization (PAN), multi-scale feature extraction with prior integration, and geometric-semantic rectification attention (GSRA).
- **Section 3: Cross-Dataset Generalization** – Evaluation of robustness across diverse lighting conditions through cross-dataset experiments (Ambient6K $\leftrightarrow$ ISTD), demonstrating PhaSR's superior generalization from single-source outdoor shadows to multi-source indoor ambient lighting.
- **Section 4: Additional Visual Comparisons** – Extensive qualitative results on ISTD+, WSRD+, INS, and Ambient6K datasets, demonstrating PhaSR's effectiveness across diverse shadow removal scenarios.
- **Section 5: Additional Feature Map Comparison** – Intermediate feature map visualization comparing PhaSR with OmniSR and DenseSR, validating the effectiveness of physically aligned prior propagation.
- **Section 6: Failure Case Study** – Analysis of challenging scenarios including dark intrinsic materials and specular surfaces, discussing limitations and future directions.
- **Section 7: Network Architecture Details** – Complete architecture specification with layer-by-layer breakdown of input/output dimensions and operations.

1. Data Loading and Preprocessing

PhaSR requires four inputs: (1) RGB image, (2) depth map, (3) normal map, and (4) semantic feature map. Depth and semantic features are extracted using pretrained DepthAnything-v2 [8] and DINO-v2 [4] models, following common practice in recent shadow removal literature [3, 7]. Normal maps are derived from depth using standard geometric conversion.

Depth-to-Normal Conversion. Given depth map $\mathbf{D} \in \mathbb{R}^{H\times W}$ and camera field-of-view (FOV = 60°), we first compute camera intrinsics:

$$f = \frac{W}{2\tan(\text{FOV}_{\text{rad}}/2)}, \quad c_x = \frac{W-1}{2}, \quad c_y = \frac{H-1}{2}, \tag{1}$$

where $\text{FOV}_{\text{rad}} = \text{FOV}_{\text{deg}} \times \pi/180$. Each pixel (x, y) with depth $z = \mathbf{D}[y, x]$ is unprojected to 3D coordinates via the pinhole camera model:

$$x_{\text{3d}} = \frac{(x - c_x)z}{f}, \quad y_{\text{3d}} = \frac{(y - c_y)z}{f}. \tag{2}$$

The resulting 3D point cloud is then converted to surface normals via spatial gradients, yielding $\mathbf{N} \in \mathbb{R}^{H\times W\times 3}$.

Normal Map Normalization. Raw normal maps $\mathbf{n}_{\text{raw}} \in [0, 1]^3$ from depth estimation are rescaled to $[-1, 1]$ and ℓ_2-normalized:

$$\mathbf{n}_{\text{rescaled}} = 2\mathbf{n}_{\text{raw}} - 1, \quad \mathbf{n}_{\text{normalized}} = \frac{\mathbf{n}_{\text{rescaled}}}{\|\mathbf{n}_{\text{rescaled}}\|_2 + \epsilon}, \tag{3}$$

where $\epsilon = 10^{-20}$ ensures numerical stability. This produces unit-length normal vectors suitable for geometric feature extraction in GSRA.

2. Algorithm Description

We provide a detailed algorithmic description of PhaSR in Algorithm 1, which illustrates the complete training pipeline including physically aligned normalization, multi-scale feature extraction with prior integration, and geometric-semantic rectification attention.

3. Cross-Dataset Generalization

To evaluate robustness across diverse lighting conditions, we conduct cross-dataset experiments where models trained on one dataset are directly tested on another without fine-tuning. As shown in Table 1, PhaSR demonstrates competitive generalization capability in both directions.

Ambient6K $\rightarrow$ ISTD. When trained on complex multi-source indoor lighting and tested on single-light outdoor shadows, PhaSR consistently outperforms both OmniSR [7] and ShadowFormer [1], achieving improvements of +1.46 dB and +3.32 dB in PSNR respectively. These results suggest that our physically aligned design—global illumination normalization via PAN and local geometric-semantic rectification via GSRA—may contribute to effective generalization from complex to simpler lighting scenarios.

ISTD $\rightarrow$ Ambient6K. The reverse direction poses greater challenges, as models trained on direct single-light shadows must adapt to multi-source ambient illumination

Algorithm 1 PhaSR Training Algorithm

Require: Shadow image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, ground truth $\mathbf{I}_{\mathrm{GT}}$
Ensure: Predicted shadow-free image $\hat{\mathbf{I}}$

Stage 1: Physically Aligned Normalization (PAN)
Gray-world: $\mathbf{I}_{\mathrm{norm}} = \mathbf{I} \cdot \frac{\mathbb{E}[\mathbf{I}]}{\mathbb{E}_c[\mathbf{I}]+\epsilon}$
Log-domain: $\log \hat{\mathbf{S}} = \mathbb{E}_{H,W}[\log(\mathbf{I}_{\mathrm{norm}} + \epsilon)]$, $\log \hat{\mathbf{R}} = \log(\mathbf{I}_{\mathrm{norm}} + \epsilon) - \log \hat{\mathbf{S}}$
Recombine: $\hat{\mathbf{I}} = \frac{\hat{\mathbf{R}}\otimes\hat{\mathbf{S}} - \min(\hat{\mathbf{R}}\otimes\hat{\mathbf{S}})}{\max(\hat{\mathbf{R}}\otimes\hat{\mathbf{S}}) - \min(\hat{\mathbf{R}}\otimes\hat{\mathbf{S}}) + \epsilon}$
  where $\hat{\mathbf{R}} = \exp(\log \hat{\mathbf{R}})$, $\hat{\mathbf{S}} = \exp(\log \hat{\mathbf{S}})$
Stage 2: Prior Extraction
Extract features: $\mathbf{F}_{\mathrm{D}}^{(i)} = \mathrm{DINOV2}(\mathbf{I})$ for $i = 0, 1, 2, 3$
Extract depth and normals: $\mathbf{D} = \mathrm{DepthV2}(\mathbf{I})$, $\mathbf{N} = \nabla\mathbf{D}$
Stage 3: Encoder with Prior Fusion
Input projection: $\mathbf{y}_0 = \mathrm{InputProj}([\hat{\mathbf{I}}, \mathbf{D}_z])$
for $\ell = 0, \ldots, 3$ **do**
  Project DINO: $\mathbf{F}_{\mathrm{d}}^{(\ell)} = \mathrm{Proj}(\mathrm{Up}(\mathbf{F}_{\mathrm{D}}^{(\ell)}))$
  Fuse: $\mathbf{y}_\ell = \mathbf{y}_\ell + \alpha_\ell \mathbf{F}_{\mathrm{d}}^{(\ell)}$
  if $\ell < 3$ **then**
    Encode: $\mathbf{c}_\ell = \mathrm{TEB}_\ell(\mathbf{y}_\ell, \mathbf{F}_{\mathrm{D}}^{(\ell)}, \mathbf{D}^{(\ell)}, \mathbf{N}^{(\ell)})$
    Downsample: $\mathbf{y}_{\ell+1} = \mathrm{Down}(\mathbf{c}_\ell)$
  end if
end for
Stage 4: Bottleneck
Concatenate scales: $\mathbf{F}_{\mathrm{cat}} = \mathrm{Conv}([\mathbf{F}_{\mathrm{D}}^{(0)}, \mathbf{F}_{\mathrm{D}}^{(1)}, \mathbf{F}_{\mathrm{D}}^{(2)}, \mathbf{F}_{\mathrm{D}}^{(3)}])$
Bottleneck: $\mathbf{c}_3 = \mathrm{PATB}([\mathbf{y}_3 + \alpha_3 \mathbf{F}_{\mathrm{d}}^{(3)}, \mathbf{F}_{\mathrm{cat}}], \mathbf{F}_{\mathrm{D}}^{(3)}, \mathbf{D}^{(3)}, \mathbf{N}^{(3)})$
Stage 5: Decoder with GSRA
for $\ell = 2, 1, 0$ **do**
  Upsample and skip: $\mathbf{u}_\ell = [\mathrm{Up}(\mathbf{c}_{\ell+1}), \mathbf{c}_\ell]$
  Feature mixing: $\mathbf{F}'_{\mathrm{g}} = \mathbf{u}_\ell + \alpha_{\mathrm{g}} \mathbf{F}_{\mathrm{g}}^{(\ell)}$, $\mathbf{F}'_{\mathrm{s}} = \mathbf{u}_\ell + \alpha_{\mathrm{s}} \mathbf{F}_{\mathrm{s}}^{(\ell)}$
  Generate KV: $\mathbf{K}_{\mathrm{g}}, \mathbf{V}_{\mathrm{g}} = \mathcal{F}_{\mathrm{g}}(\mathbf{F}'_{\mathrm{g}})$; $\mathbf{K}_{\mathrm{s}}, \mathbf{V}_{\mathrm{s}} = \mathcal{F}_{\mathrm{s}}(\mathbf{F}'_{\mathrm{s}})$
  Compute attention: $\mathbf{A}_{\mathrm{g}} = \mathrm{SoftMax}(\mathbf{Q}\mathbf{K}_{\mathrm{g}}^\top/\sqrt{d} + \mathbf{B})$
    $\mathbf{A}_{\mathrm{s}} = \mathrm{SoftMax}(\mathbf{Q}\mathbf{K}_{\mathrm{s}}^\top/\sqrt{d} + \mathbf{B})$
  Rectify: $\mathbf{A}_{\mathrm{r}} = \mathbf{A}_{\mathrm{s}} - \lambda^{(\ell)} \mathbf{A}_{\mathrm{g}}$
  Aggregate: $\mathbf{F}_{\mathrm{o}} = [\mathbf{A}_{\mathrm{r}}\mathbf{V}_{\mathrm{g}}, \mathbf{A}_{\mathrm{r}}\mathbf{V}_{\mathrm{s}}]$
  Decode: $\mathbf{c}_\ell = \mathrm{TDB}_\ell(\mathbf{F}_{\mathrm{o}}, \mathbf{F}_{\mathrm{D}}^{(\ell)}, \mathbf{D}^{(\ell)}, \mathbf{N}^{(\ell)})$
end for
Stage 6: Output and Loss
Output: $\hat{\mathbf{I}} = \mathrm{OutProj}(\mathbf{c}_0) + \mathbf{I}$
Loss: $\mathcal{L} = \lambda_{\mathrm{C}} \sqrt{\|\hat{\mathbf{I}} - \mathbf{I}_{\mathrm{GT}}\|_2^2 + \epsilon^2} + \lambda_{\mathrm{S}}(1 - \mathrm{SSIM}(\hat{\mathbf{I}}, \mathbf{I}_{\mathrm{GT}}))$

with overlapping light contributions and chromatic shifts. PhaSR maintains strong performance, outperforming competing methods by +2.33 dB over OmniSR and +4.90 dB over ShadowFormer. Notably, while all methods experience performance drops compared to in-domain training, PhaSR exhibits relatively smaller degradation, suggesting that explicit physical alignment may be associated with more robust feature learning across illumination distributions.

These results indicate that PhaSR's dual-level alignment strategy—closed-form illumination correction followed by cross-modal prior rectification—provides a design that generalizes effectively across datasets, from single-source outdoor shadows to multi-source indoor ambient lighting.

Table 1. **Cross-dataset generalization evaluation.** Models trained on one dataset and tested on another to evaluate robustness across different lighting conditions.

Method	Ambient6K → ISTD		ISTD → Ambient6K	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
ShadowFormer [1]	24.32	0.872	16.25	0.671
OmniSR [7]	26.18	0.901	18.82	0.733
PhaSR (Ours)	**27.64**	**0.923**	**21.15**	**0.798**
Reference: In-domain performance				
ShadowFormer (ISTD)	29.90	0.960	—	—
OmniSR (ISTD)	30.45	0.964	—	—
PhaSR (ISTD)	**30.73**	**0.960**	—	—
ShadowFormer (Ambient6K)	—	—	19.02	0.750
OmniSR (Ambient6K)	—	—	23.01	0.830
PhaSR (Ambient6K)	—	—	**23.32**	**0.834**

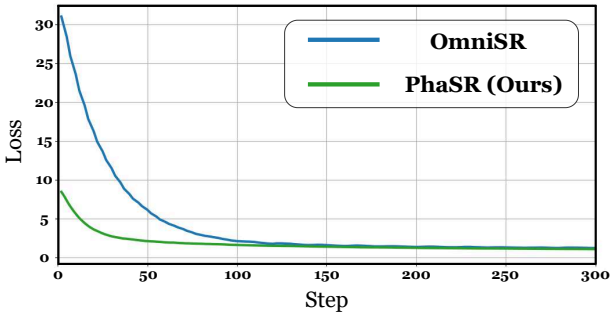

Figure 1. **Training error of OmniSR [7] with and our method on WSRD+ dataset [5].** PhaSR yields an accelerated rate of error reduction.

4. Additional Visual Comparisons

We provide additional qualitative results to demonstrate the effectiveness of PhaSR across diverse shadow removal scenarios. Figures 4, 5, 6, and 7 show comprehensive comparisons with state-of-the-art methods on ISTD+ [2], WSRD+ [5], INS [7], and Ambient6K [6] datasets, respectively.

As shown in Figure 4, PhaSR generally recovers sharper shadow boundaries and preserves texture details compared to competing methods on real-world outdoor scenes. The proposed PAN effectively normalizes illumination variations, while GSRA resolves geometric-semantic ambiguities, leading to cleaner shadow-free results.

Figure 5 demonstrates PhaSR's strong performance on high-resolution indoor scenes with complex single-source lighting. Compared to OmniSR [7] and DenseSR [3], which show some smoothing or color artifacts in certain regions, our method maintains photorealistic appearance while effectively removing shadow artifacts.

In Figure 6, we observe that PhaSR performs well on challenging synthesized indoor scenarios with indirect lighting and soft shadows. The physically aligned normalization appears to facilitate robust generalization across diverse illumination conditions, while the cross-modal atten-

tion mechanism effectively disentangles reflectance from shading.

Figure 7 further validates PhaSR's generalization capability on the challenging Ambient6K dataset, which features complex multi-source illumination and diffuse indirect lighting that goes beyond conventional shadow removal. Our method outperforms both dedicated ambient light normalization methods (IFBlend [6]) and shadow removal methods (OmniSR [7], DenseSR [3]). These results are consistent with the hypothesis that physically aligned design may facilitate handling diverse real-world lighting conditions.

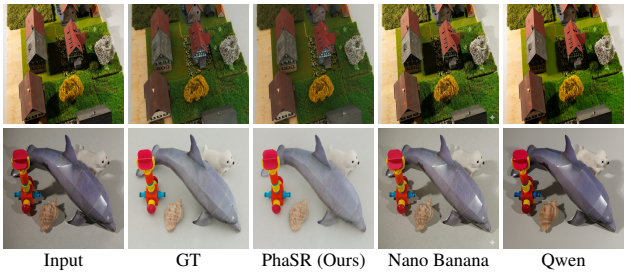

Figure 2. **Comparison between PhaSR and Multimodal LLMs.** We select two sample images from Ambient6K, PhaSR (29.41dB, 0.2s) significantly outperforms Qwen-image-edit (15.43dB, 25m) and Nano Banana (16.86dB, 30s) in both fidelity and speed (both prompted with "Remove shadows from the image").

Comparison with Multimodal LLMs. As shown in Fig. 2, despite their general image editing capabilities, MLLMs fall substantially short on shadow removal. When prompted with "Remove shadows from the image", Qwen-image-edit and Nano Banana achieve only 15.43 dB and 16.86 dB on Ambient6K, respectively, while PhaSR reaches 29.41 dB — a margin of over 12 dB. Furthermore, PhaSR completes inference in 0.2 seconds, compared to 30 seconds and 25 minutes for the MLLM counterparts. These results suggest that general-purpose MLLMs lack the task-specific physical priors necessary for precise shadow removal, whereas PhaSR's geometric-semantic alignment enables both superior fidelity and practical efficiency.

5. Additional Feature Map Comparison

Figure 8 visualizes intermediate feature maps from the encoder and decoder stages across different methods. Compared to OmniSR [7] and DenseSR [3], PhaSR's feature maps suggest several potential advantages:

- **Shadow localization:** The bottleneck features show more focused activations in shadow regions, even under complex ambient lighting.
- **Prior propagation:** Geometric and semantic information appears well-preserved through skip connections via GSRA.
- **Decoder activations:** The decoder shows progressive refinement with reduced high-frequency noise.

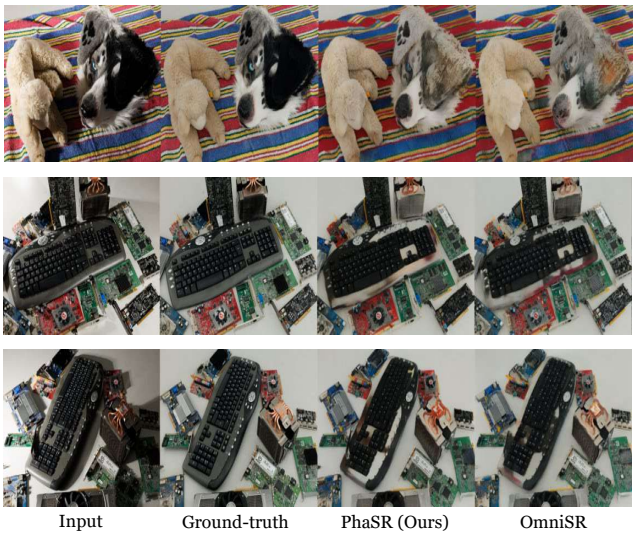

Figure 3. **Failure cases on Ambient6K [6].** Both PhaSR and existing methods struggle with shadows on intrinsically dark objects (top) or specular/metallic surfaces (bottom).

These visualizations provide qualitative evidence that the proposed physically aligned design may enable more coherent multi-scale feature learning for shadow removal.

6. Failure Case Study

Despite competitive performance across datasets, certain scenarios remain challenging for current shadow removal methods. As shown in Figure 3, both PhaSR and state-of-the-art approaches like OmniSR [7] encounter difficulties in two cases:

Dark intrinsic materials. Shadows on low-reflectance objects (e.g., black surfaces) create ambiguity between shadow-induced darkness and intrinsic material properties. Without additional cues like polarization, methods struggle to distinguish these cases, leading to under-correction or over-brightening.

Specular surfaces. Metallic and specular materials violate Lambertian assumptions underlying most shadow removal methods. View-dependent highlights and non-linear light transport cause color artifacts and inconsistent restoration when shadows interact with such surfaces.

These challenges suggest future directions including material-aware priors and non-Lambertian reflectance modeling for ambient light normalization.

7. Network Architecture Details

We provide the complete architecture specification of PhaSR in Table 2. The network consists of six main stages: physically aligned normalization, prior extraction, multi-scale encoder with prior fusion, bottleneck, hierarchical decoder with GSRA, and output generation.

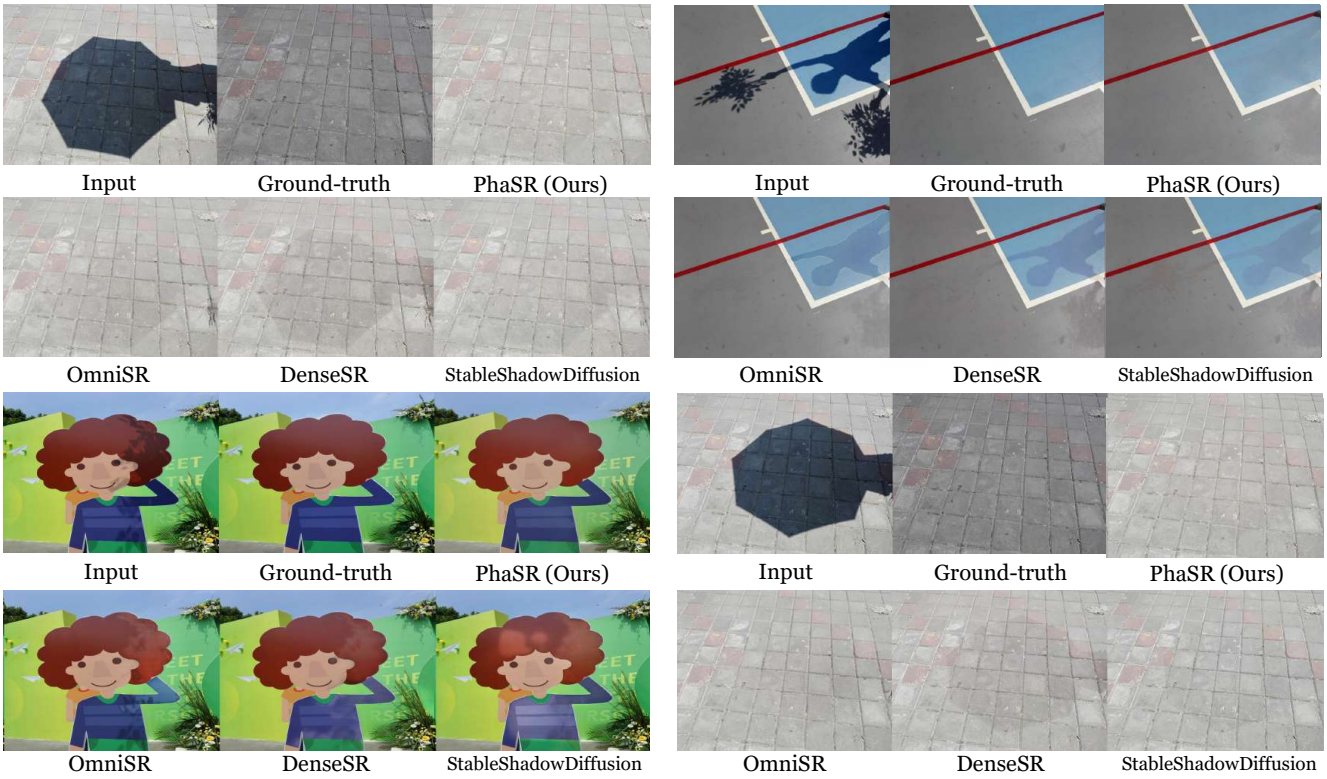

Figure 4. **Additional visual comparisons on ISTD+ [2].** PhaSR achieves superior shadow removal with sharper boundaries and better texture preservation compared to state-of-the-art methods.

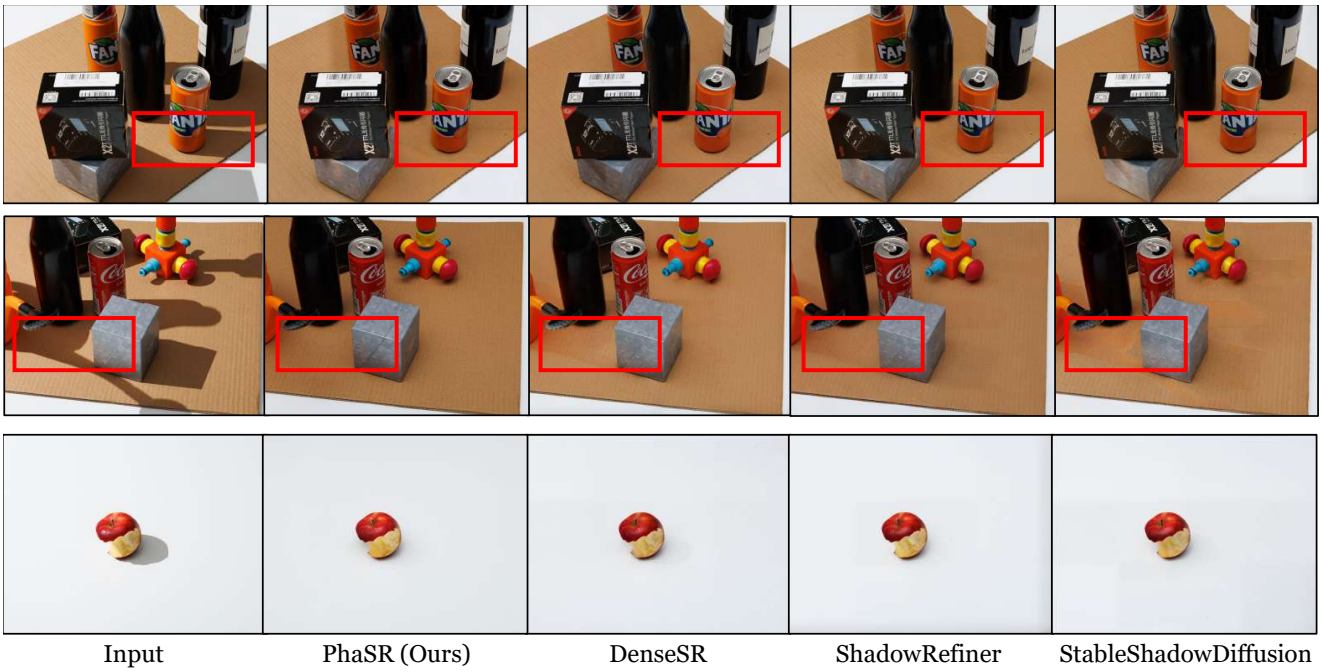

Figure 5. **Additional visual comparisons on WSRD+ [5].** Our method effectively handles high-resolution indoor scenes with complex single-source lighting while maintaining photorealistic quality.

References

[1] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: global context helps shadow removal. In AAAI, 2023. 1 and 2

[2] Hieu Le and Dimitris Samaras. Shadow removal via shadow

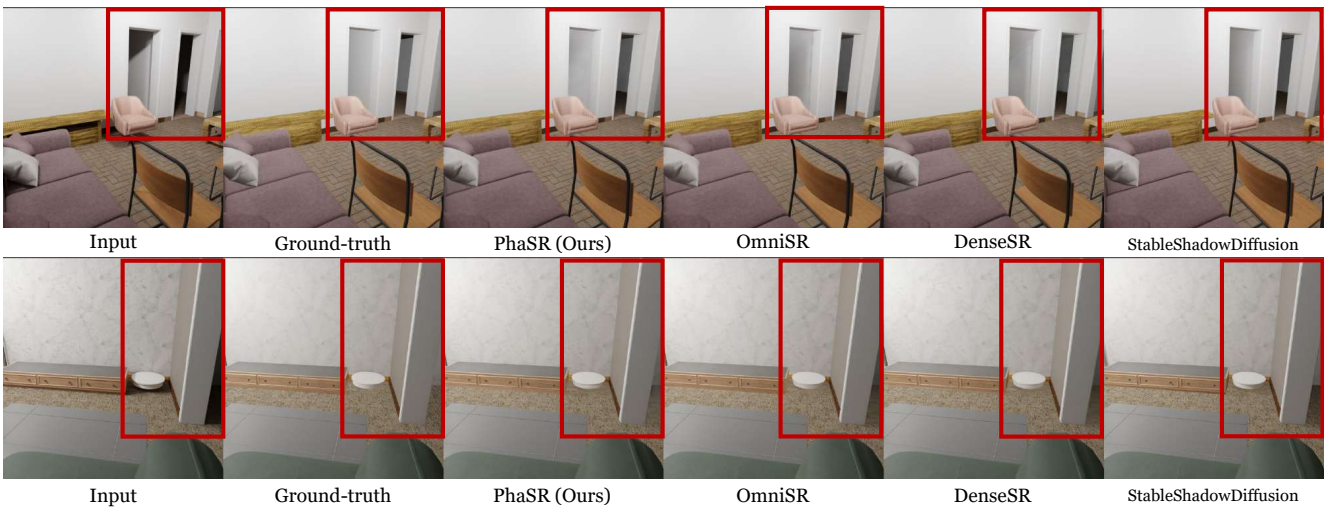

Figure 6. **Additional visual comparisons on INS [7].** PhaSR demonstrates robust generalization to synthesized indoor scenes with indirect illumination and soft shadows.

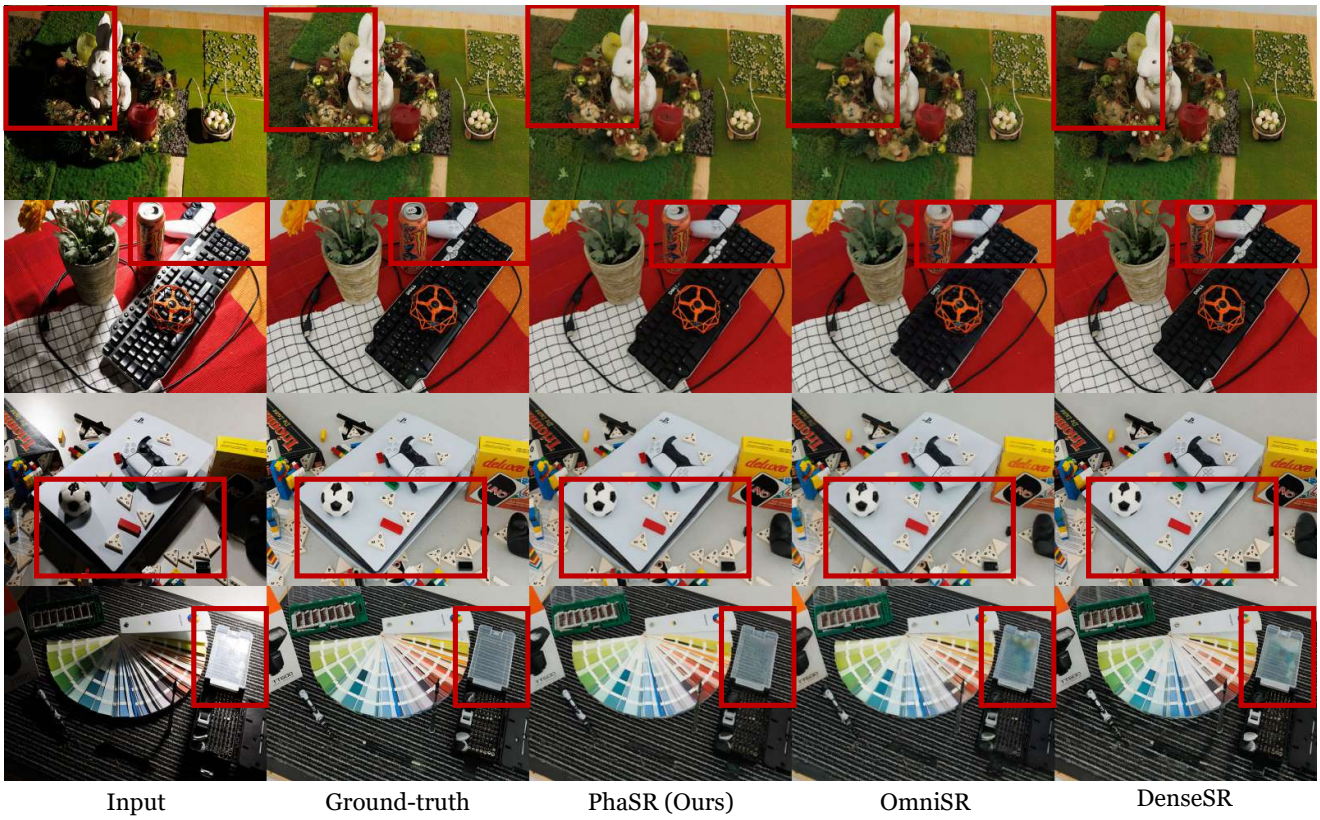

Figure 7. **Additional visual comparisons on Ambient6K [6].** PhaSR shows superior generalization to complex multi-source illumination and diffuse indirect lighting beyond conventional shadow removal, outperforming both ambient light normalization and shadow removal methods.

image decomposition. In ICCV, 2019. 2, 4, and 6

[3] Yu-Fan Lin, Chia-Ming Lee, and Chih-Chung Hsu. Densesr: Image shadow removal as dense prediction. In ACM MM, 2025. 1, 2, 3, and 6

[4] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou,

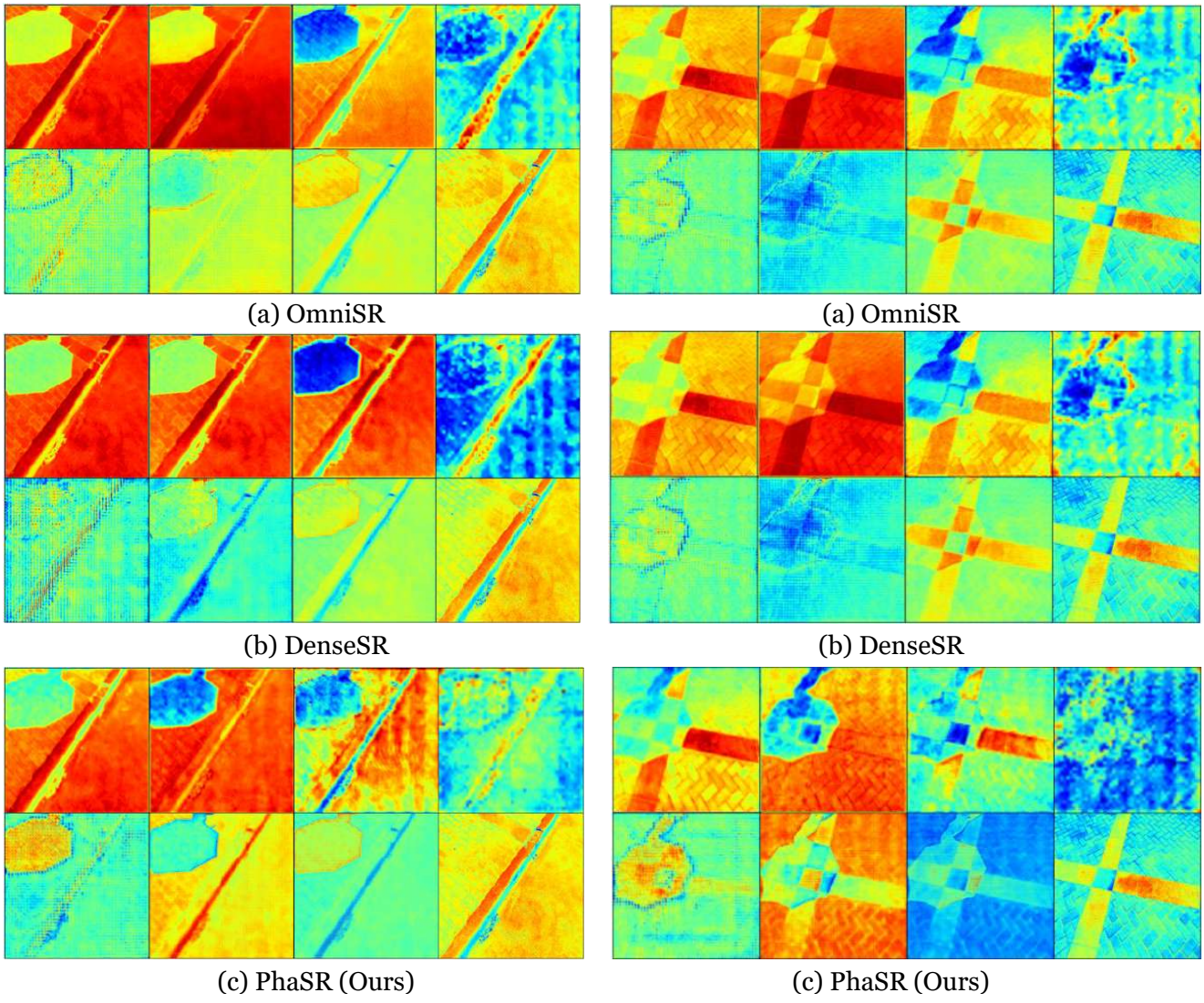

Figure 8. **Intermediate feature map visualization on ISTD+ [2].** Our method shows stronger shadow localization in bottleneck features and cleaner decoder activations compared to OmniSR [7] and DenseSR [3], validating the effectiveness of physically aligned prior propagation.

Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 1

[5] Florin-Alexandru Vasluianu, Tim Seizinger, and Radu Timofte. Wsrd: A novel benchmark for high resolution image shadow removal. In CVPRW, 2023. 2 and 4

[6] Florin-Alexandru Vasluianu, Tim Seizinger, Zongwei Wu, Rakesh Ranjan, and Radu Timofte. Towards image ambient lighting normalization. In ECCV. Springer, 2024. 2, 3, and 5

[7] Jiamin Xu, Zelong Li, Yuxin Zheng, Chenyu Huang, Renshu Gu, Weiwei Xu, and Gang Xu. Omnisr: Shadow removal under direct and indirect lighting. In AAAI, 2025. 1, 2, 3, 5, and 6

[8] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. NeurIPS, 37, 2024. 1

Table 2. **Architecture of PhaSR.** The model takes a $H \times W$ input image and processes it through PAN normalization, multi-scale Transformer encoder-decoder with DINO-V2 semantic priors and depth-derived geometric priors.

Block Name	**Output Size**	**Operation**	**Stage**
Stage 1: Physically Aligned Normalization (PAN)			
Global Estimation	$H \times W \times 3$	$\mathbf{I}_{\text{norm}} = \mathbf{I}/(\mathbb{E}[\mathbf{I}] + \epsilon)$	Eq. 2
Local Normalization	$H \times W \times 3$	$\mathbf{G}(x) = \mathbb{E}[\mathbf{I}]/(\mathbb{E}_{\Omega(x)}[\mathbf{I}] + \epsilon)$	Eq. 3
Log-domain Decomposition	$H \times W \times 3$	$\log \hat{\mathbf{S}}, \log \hat{\mathbf{R}}$ separation	Eq. 4-5
Recombination	$H \times W \times 3$	$\hat{\mathbf{I}} = \text{clamp}(\hat{\mathbf{R}} \otimes \hat{\mathbf{S}}, 0, 1)$	Eq. 5
Stage 2: Prior Extraction			
Semantic Prior (DINO-V2)			
DINO Scale 0	$H/1 \times W/1 \times 1024$	Frozen pretrained features	$\mathbf{F}_{\text{D}}^{(0)}$
DINO Scale 1	$H/2 \times W/2 \times 1024$	Frozen pretrained features	$\mathbf{F}_{\text{D}}^{(1)}$
DINO Scale 2	$H/4 \times W/4 \times 1024$	Frozen pretrained features	$\mathbf{F}_{\text{D}}^{(2)}$
DINO Scale 3	$H/8 \times W/8 \times 1024$	Frozen pretrained features	$\mathbf{F}_{\text{D}}^{(3)}$
Geometric Prior			
Depth Extraction	$H \times W \times 1$	DepthAnything-V2	$\mathbf{D}$
Normal Computation	$H \times W \times 3$	Gradient-based $\nabla\mathbf{D}$	$\mathbf{N}$
Stage 3: Multi-Scale Encoder with Prior Fusion			
Input Projection	$H \times W \times C$	Conv $4 \to C$, $C = 32$	$\mathbf{y}_0$
Encoder Level 0 ($H \times W$)			
DINO Projection	$H \times W \times C$	$\text{Conv}_{1\times1}$: $1024 \to C$	α_0
TEB (CA+DWT) $\times N_1$	$H \times W \times C$	$N_1 = 2$ layers	$\mathbf{c}_0$
Downsample	$H/2 \times W/2 \times 2C$	$\text{Conv}_{4\times4}$, stride=2	–
Encoder Level 1 ($H/2 \times W/2$)			
DINO Projection	$H/2 \times W/2 \times 2C$	$\text{Conv}_{1\times1}$: $1024 \to 2C$	α_1
TEB (CA+DWT) $\times N_2$	$H/2 \times W/2 \times 2C$	$N_2 = 2$ layers	$\mathbf{c}_1$
Downsample	$H/4 \times W/4 \times 4C$	$\text{Conv}_{4\times4}$, stride=2	–
Encoder Level 2 ($H/4 \times W/4$)			
DINO Projection	$H/4 \times W/4 \times 4C$	$\text{Conv}_{1\times1}$: $1024 \to 4C$	α_2
TEB (GSRA) $\times N_3$	$H/4 \times W/4 \times 4C$	$N_3 = 2$ layers	$\mathbf{c}_2$
Downsample	$H/8 \times W/8 \times 8C$	$\text{Conv}_{4\times4}$, stride=2	–
Stage 4: Bottleneck ($H/8 \times W/8$)			
Multi-Scale DINO Fusion	$H/8 \times W/8 \times 8C$	Concat + $\text{Conv}_{1\times1}$: $4096 \to 8C$	$\mathbf{F}_{\text{cat}}$
DINO Projection Level 3	$H/8 \times W/8 \times 8C$	$\text{Conv}_{1\times1}$: $1024 \to 8C$	α_3
PATB (GSRA) $\times N_4$	$H/8 \times W/8 \times 16C$	$N_4 = 2$ layers, concat input	$\mathbf{c}_3$
Stage 5: Hierarchical Decoder with GSRA			
Decoder Level 2 ($H/4 \times W/4$)			
Upsample	$H/4 \times W/4 \times 4C$	$\text{ConvTranspose}_{2\times2}$, stride=2	–
Skip Connection	$H/4 \times W/4 \times 8C$	Concat with $\mathbf{c}_2$	$\mathbf{u}_2$
GSRA (Sec. 3.2)	$H/4 \times W/4 \times 8C$	Geometric-Semantic Rectification	Eq. 6-10
TDB (CA+DWT) $\times N_5$	$H/4 \times W/4 \times 8C$	$N_5 = 2$ layers	$\mathbf{c}'_2$
Decoder Level 1 ($H/2 \times W/2$)			
Upsample	$H/2 \times W/2 \times 2C$	$\text{ConvTranspose}_{2\times2}$, stride=2	–
Skip Connection	$H/2 \times W/2 \times 4C$	Concat with $\mathbf{c}_1$	$\mathbf{u}_1$
GSRA (Sec. 3.2)	$H/2 \times W/2 \times 4C$	Geometric-Semantic Rectification	Eq. 6-10
TDB (CA+DWT) $\times N_6$	$H/2 \times W/2 \times 4C$	$N_6 = 2$ layers	$\mathbf{c}'_1$
Decoder Level 0 ($H \times W$)			
Upsample	$H \times W \times C$	$\text{ConvTranspose}_{2\times2}$, stride=2	–
Skip Connection	$H \times W \times 2C$	Concat with $\mathbf{c}_0$	$\mathbf{u}_0$
GSRA (Sec. 3.2)	$H \times W \times 2C$	Geometric-Semantic Rectification	Eq. 6-10
TDB (CA+DWT) $\times N_7$	$H \times W \times 2C$	$N_7 = 2$ layers	$\mathbf{c}'_0$
Stage 6: Output Generation			
Output Projection	$H \times W \times 3$	$\text{Conv}_{3\times3}$: $2C \to 3$	–
Residual Connection	$H \times W \times 3$	$\hat{\mathbf{I}} = \text{OutProj}(\mathbf{c}'_0) + \mathbf{I}$	Final