Guided Lensless Polarization Imaging

Supplementary Material

This supplementary material provides additional details that supplement the paper *"Guided Lensless Polarization Imaging"*. Section 1 describes the optical setup used in our experiments. Section 2 outlines the implementation of the physics-based reconstruction baselines (FISTA and ADMM). Section 3 presents additional details of the RGB-guided refinement network and Section 4 summarizes further implementation details and shows complementary results.

1. Optical Setup

Our prototype lensless polarization camera, shown in Figure 2, combines a random diffuser with a polarization mask following the design of Elmalem and Giryes [6], Kraicer et al. [8], as described in the Lensless Polarization Imaging Setup section of the main paper (Section 3.1). These two optical elements jointly encode spatial and polarization information in the image plane. We use a $0.5°$ diffuser (Edmund Optics #47-860), mounted on a 12.3 MP color CMOS sensor (Thorlabs CS126CU) with a pixel pitch of $3.45\,\mu$m.

The polarization mask is fabricated by cutting a linear polarizing film (Thorlabs LPVISE2X2) into stripes approximately $880\,\mu$m wide (corresponding to $\sim$256 sensor pixels). These stripes are placed on a glass substrate and arranged into the desired polarization-angle pattern, following a repeating sequence of $0°, 45°, 90°$, and $135°$ (repeated four times), as illustrated in Figure 1. The mask is designed to support one-shot acquisition of the four polarization angles while remaining simple to fabricate under lab conditions, and is similar to prior designs [3, 6, 8]. The mask assembly is then mounted just above the sensor's cover glass to ensure accurate, high-fidelity polarization sampling at each angle. Incoming light first passes through the diffuser and is multiplexed before reaching the polarization mask, after which the encoded light is recorded by the sensor. To measure the system's polarization-independent PSF, we image a point light source after removing the camera's polarization mask. To measure the angular response introduced by the polarization mask (at $0°$, $45°$, $90°$, and $135°$), we place a broadband, uniform light source in front of the bare lensless camera (after removing the diffuser) and control the incident polarization angle using an external motorized rotating linear polarizer (Thorlabs LPVISE100-A) mounted on a Thorlabs ELL14K stage. These angular response measurements are shown in Figure 3 of the main paper. We operate in a regime where the PSF is approximately shift-invariant (LSI), which simplifies our reconstruction to a linear shift-invariant model. To obtain reference polarization images at each angle, an RGB sensor (UI-3590LE-C-HQ) is placed behind the rotating polarizer. These reference measurements are also used to compute the RGB guidance image, as defined in Equation (2) of the main paper, in the same manner as the simulated datasets. A photograph of the complete optical setup is shown in Figure 2. The raw 16-bit lensless sensor data are converted to 32-bit floating point and normalized to the $[0, 1]$ range. White balance correction is then applied by scaling the red and blue channels to match the mean intensity of the green channel, before further processing.

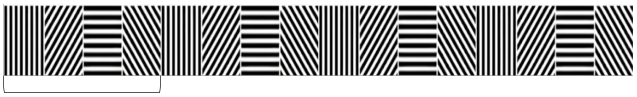


Figure 1. Illustration of our polarization mask, containing four repetitions of striped polarizers at orientations $0°, 45°, 90°$, and $135°$

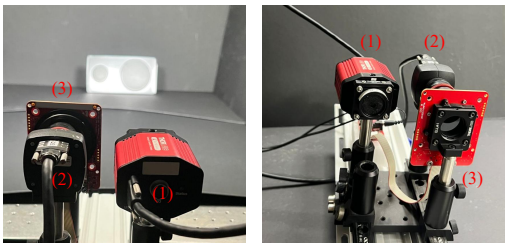


Figure 2. The experimental optical setup from two viewpoints. **Left:** Front view showing the imaging target and the two-sensor setup: (1) lensless polarization camera and (2) RGB reference camera. **Right:** Back view of the setup showing (3) the rotation stage with a mounted linear polarizer positioned in front of (2) the RGB camera for capturing reference images. The lensless camera prototype (1) consists of a diffuser and a manually assembled polarization mask mounted on the sensor.

2. Physics-Based Reconstruction Implementation Details

For both FISTA and ADMM, the polarization intensity image $\hat{\mathbf{x}}$ is recovered by solving the optimization problem in Equation (3) with its forward operator in Equation (1) of the main paper. Both solvers are GPU-accelerated using CuPy.

2.1. FISTA

We adopt the FISTA solver with Haar-based anisotropic total variation (TV) from the publicly available implementa-

tion of SpectralDiffuserCam [9].

For simulated data (PIP, UPLight, and ZJU-RGB-P), FISTA is run for 10k iterations, an empirically chosen value that ensures visual convergence and stable reconstruction quality. For real-world measurements, the iteration count is reduced to 500 in both color and grayscale, which is sufficient for convergence.

A fixed step size of $1/(L \cdot c)$ is used for FISTA's update, where L is the Lipschitz constant of the forward operator and c is a tuning factor. Following prior work, we set $c = 45$ for synthetic data [9], $c = 100$ for the PSF-mismatch ablation, and use a more conservative $c = 1000$ for real measurements to improve stability under noise and forward-model mismatch. All reconstructions are initialized with zeros.

We regularize $\mathbf{x}$ using a Haar-based anisotropic 3DTV prior across both spatial and polarization dimensions. Let λ denote the global regularization strength and λ_w the relative weighting of the polarization axis with respect to the spatial axes. The directional weights are defined as

$$w_{\text{ax}} = \begin{cases} \lambda_w, & \text{if } \texttt{axis} = \text{polarization}, \\ 1, & \text{if } \texttt{axis} = \text{spatial}. \end{cases}$$

Each FISTA iteration then performs a proximal update that combines non-negativity with Haar-based TV:

$$\mathbf{x} \leftarrow \tfrac{1}{2}\Big(\max(\mathbf{x}, 0) \; + \; \texttt{tvApproxHaar}\big(\mathbf{x}, \tfrac{\lambda}{L \cdot c}, w_{\text{ax}}\big)\Big),$$

Where the scaling factor $L \cdot c$ is applied consistently in both the gradient step and the TV thresholding.

We set $\lambda = \lambda_w = 5 \times 10^{-5}$ for synthetic data and increase both to 5×10^{-3} for real measurements, which require stronger TV regularization. All hyperparameters were selected via grid search to balance denoising and structure preservation. For real data in the three-angle grayscale configuration, the measured PSF and mask are converted to grayscale before reconstruction.

2.2. ADMM

We also solve Eq. (3) using scaled ADMM [5], with updates:

$$(A^\top A + \rho I)\, v^{t+1} = A^\top y + \rho\,(z^t - u^t), \tag{1}$$

$$z^{t+1} = \text{prox}_{(\lambda/\rho)\,\text{TV}}\big(v^{t+1} + u^t\big)\,, \tag{2}$$

$$u^{t+1} = u^t + v^{t+1} - z^{t+1}. \tag{3}$$

Here, $\mathbf{v}$ is the data-fidelity variable, $\mathbf{z}$ is the regularization variable enforcing the TV and non-negativity constraints, and $\mathbf{u}$ is the scaled dual variable (Lagrange multiplier) enforcing consensus between $\mathbf{v}$ and $\mathbf{z}$. The variables $\mathbf{A}$ and $\mathbf{y}$ represent the system's forward operator and the measurement vector, respectively, as defined in Eq. (3) in the main paper.

Unlike FISTA, in ADMM the non-negativity constraint is applied to the TV-regularized variable z, i.e., $z^{t+1} \leftarrow \max\big(z^{t+1}, 0\big)$, rather than directly to v.

The TV proximal operator uses the same Haar-based anisotropic formulation as in the FISTA solver. We solve (1) inexactly using conjugate gradients, since the forward operator $\mathbf{A}$ includes spatial masking and cropping, making $A^\top A$ non-shift-invariant and thus not diagonalizable in the Fourier domain. We use $\rho = 0.15$, $\lambda = 3 \times 10^{-5}$, $\lambda_w = 6 \times 10^{-5}$, 200 ADMM iterations, and CG tolerance 10^{-3} with 30 inner iterations for all simulation datasets. We use ADMM only for simulated datasets in our ablations; all real-data reconstructions are obtained with FISTA.

Note on RGB Data. For both the FISTA and ADMM implementations, when processing four-angle RGB measurements, the 3D total-variation regularization (spatial and polarization dimensions) is applied separately to each color channel.

3. RGB-Guided Deep Refinement Implementation Details

The second stage of our reconstruction pipeline utilizes an RGB-guided refinement network based on SwinFuSR [2], as described in the RGB-Guided Deep Refinement section of the main paper (Section 3.4).

The network architecture follows the structure of SwinFuSR and is composed of three core modules: **Extraction**, **Fusion**, and **Reconstruction**.

The Extraction module applies shallow convolutions and Swin Transformer Layers (STLs) to encode features from the initial reconstruction and RGB image. The Fusion module integrates these feature streams through Attention-guided Cross-domain Fusion (ACF) blocks. The Reconstruction module refines the fused representation with additional STLs and convolutions to produce the final output with a skip connection for the initial reconstruction.

The module depths are configured as follows: two STLs in the Extraction module, three ACF blocks in the Fusion module, and three STLs in the Reconstruction module, consistent with the original SwinFuSR configuration.

The original implementation of SwinFuSR was designed for guided thermal Super-Resolution (SR), where the low-resolution (LR) thermal input image is upsampled to high-resolution (HR) before entering the network. We omit this upsampling step because both our initial reconstruction and the RGB guidance image are already at the target high resolution, making the operation unnecessary. We used a batch size of 1 due to memory constraints.

Table 1. Baseline comparison under the three-angle grayscale configuration on the UPLight and ZJU-RGB-P evaluation datasets.

Model	PSNR↑	SSIM↑	LPIPS↓
UPLight			
FlatNet	10.78	0.27	0.98
PolarAnything (FISTA input)	11.84	0.36	0.98
PolarAnything (RGB input)	11.98	0.40	0.93
FISTA	16.72	0.26	0.53
FISTA + Transf.	17.93	0.44	0.53
Ours (FISTA input)	**20.49**	**0.52**	**0.32**
ZJU-RGB-P			
FlatNet	16.73	0.54	0.57
PolarAnything (FISTA input)	19.05	0.58	0.42
PolarAnything (RGB input)	19.96	0.62	0.38
FISTA	14.50	0.46	0.44
FISTA + Transf.	27.20	0.89	0.19
Ours (FISTA input)	**31.19**	**0.97**	**0.07**

4. Additional Experimental Results

This section details complementary information for the experimental results reported in Section 4 of the main paper.

4.1. Additional Implementation Details

Training Dataset We selected the Polarimetric Imaging for Perception (PIP) dataset [4] because it is the largest publicly available dataset providing aligned RGB and polarization data suitable for our task. However, the publicly released version contains only derived polarization metrics (Angle of Linear Polarization (AoLP) and Degree of Linear Polarization (DoLP)) along with RGB images, which are computed from the raw polarization intensity images ($I_0, I_{45}, I_{90}, I_{135}$). Since our method requires these raw intensity measurements, we obtained them directly from the dataset authors upon request, along with their established preprocessing pipeline. We split the data into 8,538 training, 2,717 validation, and 1,372 test images without scene overlap.

Evaluation Metrics We used three metrics for quantitative evaluation: PSNR, SSIM, and LPIPS (VGG). For the three-angle grayscale polarization intensity images configuration, PSNR and SSIM are averaged over channels, while LPIPS is computed per channel (converted to RGB) and averaged. For the 12-channel RGB four-angle configuration, PSNR and SSIM are averaged over all 12 channels. LPIPS is computed per RGB triplet (i.e., three channels representing a single polarization angle) and then averaged over the four triplets.

4.2. Synthetic Data Results

This section further elaborates on the synthetic data results presented in Section 4.1 of the main paper.

FlatNet and PolarAnything FlatNet [7] and PolarAnything [10] are two additional baselines included in our comparisons, as detailed in the Experimental Results section of the main paper (Section 4.1).

FlatNet provides both separable and non-separable variants depending on the structure of the PSF. We use the non-separable model, which aligns with the characteristics of our measured PSF. Because its U-Net architecture expects input dimensions divisible by 32, we pad the 250×250 sensor images to the nearest valid size (256×256). Training was performed for 50 epochs (approximately the same number of iterations in the paper). The rest of the parameters were the same as their code's base config and only MSE loss, which yielded the best results.

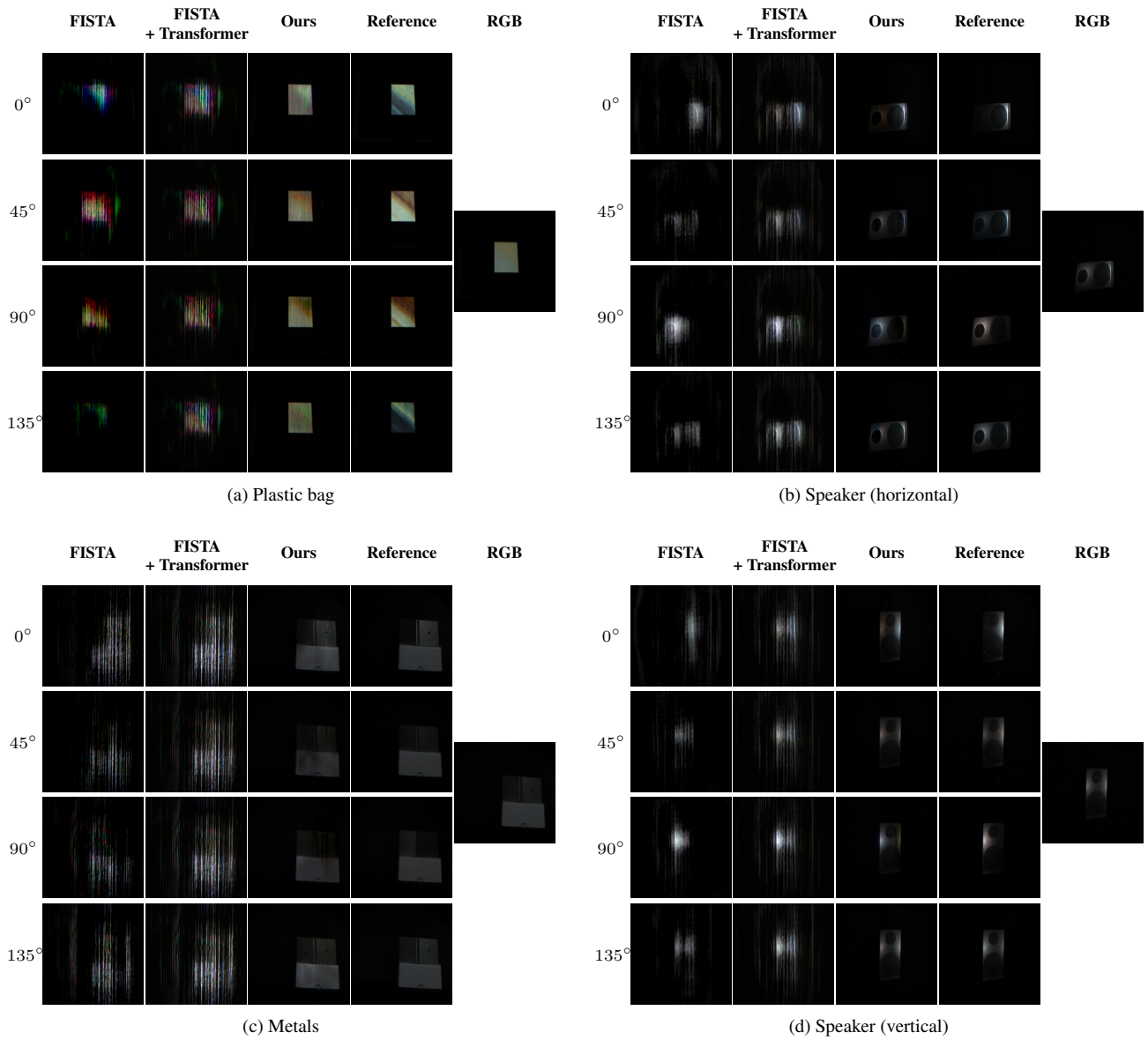
For PolarAnything, we likewise pad the 250×250 inputs to 256×256 to satisfy the spatial-resolution requirements of the diffusion U-Net. We train the network for each configuration (RGB / FISTA input) for 20 epochs, which is sufficient for convergence; all other training parameters follow those reported in the paper. PolarAnything was trained on two NVIDIA RTX A6000 GPUs with a batch size of 32.

For both baselines, the reconstructed images are cropped to their original resolution before computing all evaluation metrics. Table 1 presents the generalization performance of FlatNet and PolarAnything on the supplementary UPLight and ZJU-RGB-P datasets, which were not featured in the main paper (see Table 2 in the main paper). Both models show limited generalization, notably on the UPLight dataset, where their results are worse than the FISTA baseline, clearly demonstrating a significant performance gap compared to our proposed approach and the unguided FISTA+Transformer baseline.

Fine-Tuning Evaluation. Table 2 shows the performance on **ZJU-RGB-P and UPLight** before and after fine-tuning on 10 image pairs, demonstrating quantitative gains with minimal target-domain data. Table 3 summarizes the performance on **PIP** before and after this fine-tuning. As expected, domain shifts lead to degradation on the source domain (PIP), more pronounced for UPLight due to its larger domain shift (underwater scenes vs street scenes in PIP and ZJU-RGB-P). However, this degradation is moderate and represents a trade-off that enables improved performance on the target domain using only a small amount of new data.

4.3. Real-world Results

The qualitative results on real lensless polarization data in the 3-angle grayscale configuration are shown in Figure 5



(a) Plastic bag

(b) Speaker (horizontal)

(c) Metals

(d) Speaker (vertical)

Figure 3. Qualitative comparison on four real scenes: plastic bag, horizontal speaker, metals, and vertical speaker. For each scene, rows correspond to the four polarization angles, and columns show FISTA, FISTA+Transformer, our method, the reference polarization image, and the RGB guidance image used by our method (rightmost column).

in the main paper. For the same scenes, the results under the four-angle RGB configuration further confirm that our method consistently surpasses the reference approaches, yielding accurate and reliable reconstructions as seen in Figure 3.

4.4. Ablation Studies

This section provides supplementary details for the ablation studies in Section 4.3 of the main paper.

Simple Fusion While Table 3 in the main paper shows that simple feature fusion does not result in a significant quantitative performance drop compared to our method (which uses cross-attention fusion), the qualitative differences are present. As illustrated in Figure 4, simple fusion yields results with visual intensity that appears different and less accurate than our approach across both the simulated (UPLight, ZJU-RGB-P) and real-scene datasets. Furthermore, in the real-scene example of the first row, this simple fusion mechanism fails to effectively integrate the RGB and initial reconstruction features.

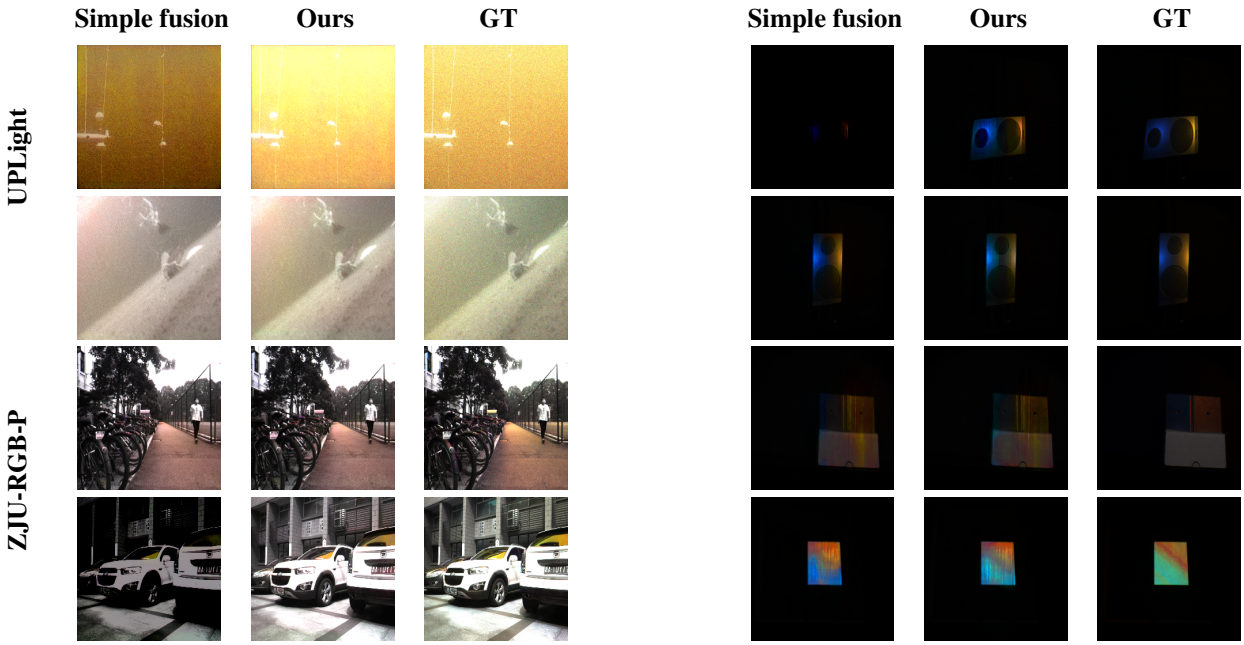


Figure 4. Qualitative comparison of **simple fusion** and **our method**. Left: simulated datasets (UPLight, ZJU-RGB-P). Right: real scenes. Each polarization triplet (0°, 45°, 90°) is visualized as an RGB composite.

Table 2. Performance before and after fine-tuning on 10 training pairs from ZJU-RGB-P or UPLight, under the **four-angle RGB** and **three-angle grayscale** configurations, using the base model with FISTA input. The reported results exclude the 10 samples used for fine-tuning.

Model	UPLight			ZJU-RGB-P		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Color						
Base	20.06	0.51	0.28	30.36	0.96	0.04
Fine-tuned	**21.99**	**0.55**	**0.18**	**31.66**	**0.97**	**0.03**
Grayscale						
Base	20.49	0.52	0.32	31.16	**0.97**	0.07
Fine-tuned	**24.67**	**0.56**	**0.23**	**32.74**	**0.97**	**0.05**

Table 3. **Performance on the PIP dataset** before and after fine-tuning on 10 training pairs from UPLight or ZJU-RGB-P, evaluated under the four-angle RGB and three-angle grayscale configurations using the base model with FISTA input.

Model	PSNR↑	SSIM↑	LPIPS↓
Color			
Base - before	33.05	0.95	0.04
FT (UPLight)	28.95	0.92	0.06
FT (ZJU-RGB-P)	31.53	0.95	0.05
Grayscale			
Base - before	35.13	0.97	0.03
FT (UPLight)	31.26	0.96	0.07
FT (ZJU-RGB-P)	33.53	0.97	0.04

PSF Mismatch The results presented in Table 3 of the main paper demonstrate our model's robust generalization to unseen optics during inference. Specifically, we used two additional Point Spread Functions (PSFs) from Antipa et al. [1] (PSF #1) and Monakhova et al. [9] (PSF #2) to simulate and reconstruct the UPLight and ZJU-RGB-P datasets. These PSFs are distinct from the training simulation and reconstruction PSF used for the PIP dataset. The visual comparison of all three PSFs is provided in Figure 5. All PSFs share a similar speckle-like structure but differ in geometry.

RGB Guidance Similar to the FISTA+Transformer baseline, we train an RGB+Transformer model to highlight the limitations of RGB-only reconstruction. The RGB-only model (Table 4) underperforms our method and fails to generalize to UPLight (an out-of-distribution (OOD) dataset), as it lacks polarization information. In contrast, our method preserves consistency with the physics-based polarization initialization via a skip connection, while RGB provides complementary high-frequency structure through cross-attention (see Figure 4 in the main paper).

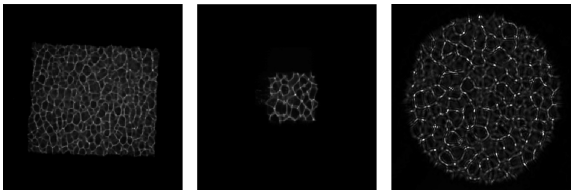


Figure 5. PSFs used in ablation study (#1 and #2) and the training PSF (measured from our system). **Left:** PSF #1. **Middle:** PSF #2. **Right:** Measured PSF used for training.

Table 4. RGB ablation under the same training pipeline as our method. FISTA+T denotes the baseline derived from our architecture, in which the Transformer refines only the FISTA reconstruction without RGB guidance (i.e., the same input is fed to both branches). RGB+T uses the same architecture with RGB-only input. UPLight is out-of-distribution (OOD) relative to the training data.

Model	**PIP**		**UPLight**	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Polar (FISTA+T)	28.85	0.88	17.93	0.44
RGB (RGB+T)	32.28	**0.97**	13.11	0.48
Ours	**35.13**	**0.97**	**20.49**	**0.52**

Translation Augmentation We evaluate robustness to misalignment by applying random translations to the RGB guidance image on simulated data (Table 5). Without augmentation, performance degrades under such shifts, whereas training with translation augmentation largely eliminates this sensitivity.

We further evaluate a wider range of translation magnitudes on simulated data and observe stable performance even beyond ± 4 pixels. However, in real-world experiments, we find that using ± 4 pixels during training provides the best overall performance, reflecting the typical level of residual misalignment in our setup. This is particularly important in practice, where perfect alignment cannot be guaranteed, motivating its use in the main paper.

Table 5. Robustness to test-time translations (total shift: 0–4 px). Metrics are reported as PSNR / SSIM, averaged over two seeds.

Dataset	**0px**	**1px**	**2px**	**3px**	**4px**
Ours (no translation augmentation, grayscale)					
UPLight	19.92 / .54	19.84 / .52	19.74 / .51	19.61 / .50	19.46 / .50
RGBP	31.74 / .97	25.84 / .94	23.65 / .91	22.56 / .90	21.94 / .89
PIP	35.53 / .97	27.91 / .94	25.36 / .92	24.05 / .91	23.23 / .90
Ours (trained with translation augmentation, grayscale)					
UPLight	20.49 / .52	20.49 / .52	20.48 / .52	20.47 / .52	20.46 / .52
RGBP	31.19 / .97	31.15 / .97	31.10 / .97	31.03 / .97	30.95 / .97
PIP	35.13 / .97	35.08 / .97	35.03 / .97	34.96 / .97	34.88 / .97

References

[1] Nick Antipa, Grace Kuo, Reinhard Heckel, Ben Mildenhall, Emrah Bostan, Ren Ng, and Laura Waller. Diffusercam: lensless single-exposure 3d imaging. *Optica*, 5(1):1–9, 2018. 5

[2] Cyprien Arnold, Philippe Jouvet, and Lama Seoud. Swinfusr: an image fusion-inspired model for RGB-guided thermal image super-resolution. In *CVPR*, pages 3027–3036, 2024. 2

[3] Nakkyu Baek, Yujin Lee, Taeyoung Kim, Jaewoo Jung, and Seung Ah Lee. Lensless polarization camera for single-shot full-Stokes imaging. *APL Photonics*, 7(11), 2022. 1

[4] Michael Baltaxe, Tomer Pe'er, and Dan Levi. Polarimetric imaging for perception. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. 3

[5] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3 (1):1–122, 2011. 2

[6] Shay Elmalem and Raja Giryes. A lensless polarization camera. In *Computational Optical Sensing and Imaging*, pages CTh7A–1. Optica Publishing Group, 2021. 1

[7] Salman Siddique Khan, Varun Sundar, Vivek Boominathan, Ashok Veeraraghavan, and Kaushik Mitra. FlatNet: Towards photorealistic scene reconstruction from lensless measurements. *IEEE TPAMI*, 44(4):1934–1948, 2020. 3

[8] Noa Kraicer, Shay Elmalem, Erez Yosef, Hani Barhum, and Raja Giryes. A lensless polarization camera. *arXiv preprint arXiv:2603.17156*, 2026. 1

[9] Kristina Monakhova, Kyrollos Yanny, Neerja Aggarwal, and Laura Waller. Spectral diffusercam: lensless snapshot hyperspectral imaging with a spectral filter array. *Optica*, 7(10): 1298–1307, 2020. 2, 5

[10] Kailong Zhang, Youwei Lyu, Heng Guo, Si Li, Zhanyu Ma, and Boxin Shi. Polaranything: Diffusion-based polarimetric image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 26466–26476, 2025. 3