INFERRING OBJECTS AROUND CORNERS FROM INDIRECT WALL LIGHT: STRUCTURE-PRESERVING SILHOUETTE RECOVERY WITH RUBIN

Anonymous authors
Paper under double-blind review

ABSTRACT

We ask whether a conventional RGB camera can reveal a target silhouette from indirect wall light around a corner. We propose RUBIN, a structure-preserving framework that addresses acquisition-related radiance variation and boundary ambiguity caused by diffuse transport. Radiance-Consistent Structural Learning (RCSL) constructs position-preserving radiance views and aligns global representations, co-located local features, and dense structural relations. Variance and covariance regularization preserve descriptor diversity and reduce redundancy, while an acquisition-factor branch discourages encoding of sampled acquisition factors. Uncertainty-Gated Boundary Region Refinement (UGBR) combines coarse region predictions with multiscale boundary evidence and weights relational corrections using region uncertainty and boundary agreement. Experiments on five measured datasets show stronger overall silhouette recovery than twenty representative methods. Ablations support the complementary roles of RCSL and UGBR, and observation-variation experiments examine the shape information retained as wall-light patterns change. The recovered foreground extent, major components, and basic contours support preliminary hidden-target shape inference from wall light in this setup.

1 INTRODUCTION

Can indirect light on a visible wall reveal the shape of an object around a corner? Light from the hidden region undergoes indirect propagation and wall reflection, forming faint, spatially mixed patterns within the camera's field of view. Corner-induced intensity variations have been used to infer hidden-person counts and motion from one-dimensional videos (Bouman et al., 2017), while PAC-Net estimates walking trajectories from wall observations (Wang et al., 2023). We ask whether wall observations can yield two-dimensional masks conveying the target outline and major-component placement. Our input is a steady-state RGB relay wall observation captured by a conventional camera. The acquisition setup in Figure 1 measures light from two-dimensional target images presented behind an occluder, allowing silhouette recovery to be evaluated against known target masks.

Two forms of structural ambiguity make this task difficult. First, exposure, channel response, saturation, wall response, noise, and blur can change the observation while the target shape remains fixed. An encoder can mistake these intensity variations for target evidence or learn redundant features in weak-response regions. The same target can therefore produce different intensity distributions, making stable spatial relationships essential for recovering its shape. Second, diffuse transport spreads boundary responses, producing ambiguous transitions, fragmented interiors, and spurious foreground activations. Because a wall pixel can mix responses from multiple target locations, foreground assignments based on local logits can remain ambiguous. Unrestricted boundary updates can also propagate wall texture into the foreground. Recovering the silhouette therefore requires determining both which regions belong to the target and where their boundaries should lie.

We couple radiance-consistent learning with uncertainty-gated refinement in RUBIN to address these ambiguities. Radiance-Consistent Structural Learning constructs position-preserving radiance views and aligns global, local, and dense representations. Variance and covariance regularization preserve descriptor diversity and reduce redundancy, while an acquisition-factor branch discourages encoding of sampled acquisition factors. Uncertainty-Gated Boundary Region Refinement uses the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

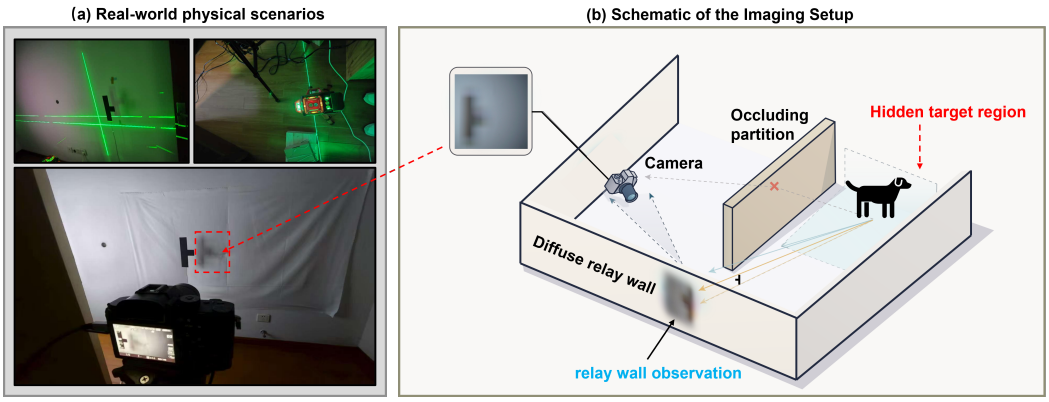

Figure 1: Overview of the non-line-of-sight acquisition setup, including the corner occlusion layout and the relay wall imaging apparatus.

resulting structural features to predict coarse regions, estimates boundary confidence from multiscale disagreement, and weights relational corrections using region uncertainty. Time-resolved NLOS systems recover hidden three-dimensional geometry using pulsed illumination and photon timing (O'Toole et al., 2018; Nam et al., 2021). Across five measured datasets, RUBIN recovers basic target outlines from conventional-camera RGB wall observations and achieves stronger overall foreground recovery than twenty representative methods.

Our main contributions are as follows:

- We demonstrate the feasibility of recovering basic hidden-target shape from indirect wall light captured by a conventional camera and introduce RUBIN for structure-preserving silhouette recovery.
- We introduce RCSL for structural consistency under modeled radiance variations and UGBR for boundary and region correction guided by uncertainty and multiscale agreement.
- Experiments on five measured datasets establish RUBIN's overall advantage over twenty representative methods. Component ablations examine the contributions of RCSL and UGBR, and observation-variation experiments evaluate the shape information retained as wall observations change.

2 Related Work

Recovering hidden scenes from indirect light. Time-resolved methods recover hidden geometry using photon timing, transport priors, sparse scanning, and temporal aggregation (O'Toole et al., 2018; Mu et al., 2025; Li et al., 2023b; Cui et al., 2024; Li et al., 2024a). Ordinary-camera methods exploit penumbrae, optimal transport, learned transport representations, or spectral separation (Saunders et al., 2019; Geng et al., 2022; Zhang et al., 2025; Hashemi et al., 2025), while related work estimates hidden-person motion and trajectories (Bouman et al., 2017; Wang et al., 2023). RUBIN focuses on binary silhouette recovery from a single RGB wall observation, preserving foreground structure under radiance variation and spatial mixing.

Learning stable structural representations. VICReg and its information-theoretic analysis motivate view consistency and statistical regularization (Bardes et al., 2022a; Shwartz-Ziv et al., 2023). Gradient reversal encourages domain-invariant features (Ganin et al., 2016), while PooDLe combines pooled and dense self-supervised objectives (Wang et al., 2025). RCSL adapts these principles to position-preserving radiance views, jointly aligning global context, co-located features, and dense relations while discouraging acquisition-factor encoding. These constraints promote stable spatial relations for silhouette recovery.

Boundary-aware and uncertainty-guided segmentation. GateNet, FAP-Net, and PGNet improve foreground prediction through gated transmission, boundary guidance, and pyramid feature

fusion (Zhao et al., 2020; Zhou et al., 2022; Xie et al., 2022). Cascade-CLIP explores multilevel vision-language alignment (Li et al., 2024b), while ZoomNeXt encourages confident predictions through uncertainty-aware regularization (Pang et al., 2024). Topology-aware uncertainty and constrained implicit representations address structural reliability and geometric constraints (Gupta et al., 2023; Schneider et al., 2024). For spatially mixed wall observations, UGBR combines coarse-region uncertainty with multiscale boundary agreement to control foreground refinement, supported by a weak auxiliary topology constraint.

3 METHOD

3.1 PROBLEM FORMULATION

We recover a hidden target's binary mask $\mathbf{y} \in \{0,1\}^{H\times W}$ in target-image coordinates from a normalized relay wall observation $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$. Let $\mathbf{x}_0$ denote the unnormalized RGB intensity, $\mathbf{u}$ a wall position, and $\mathbf{q}$ a coordinate on the hidden two-dimensional target plane. With target-plane domain Ω_{tar} and wavelength domain Λ, a transport-inspired abstraction is

$$\mathbf{x}_0(\mathbf{u}) = \int_{\Omega_{\text{tar}}} \int_{\Lambda} \kappa(\mathbf{u}, \mathbf{q}, \lambda; \boldsymbol{\xi}_{\text{geom}}) \mathbf{a}(\mathbf{q}, \lambda) \mathbf{1}_{\mathbf{y}}(\mathbf{q}) \, \mathrm{d}\lambda \, \mathrm{d}\mathbf{q} + \mathbf{b}(\mathbf{u}) + \mathbf{n}(\mathbf{u}), \tag{1}$$

$$\widehat{\mathbf{y}} = \sigma(f_\theta(\mathbf{x})) \, . \tag{2}$$

Here, κ describes indirect transport under geometry $\boldsymbol{\xi}_{\text{geom}}$; $\mathbf{a}$ gives the foreground spectral radiance, and $\mathbf{1}_{\mathbf{y}}$ selects its support. The term $\mathbf{b}$ includes the relay response from the background and ambient illumination, while $\mathbf{n}$ represents measurement noise. Transport and appearance remain latent: RUBIN learns the mapping from $\mathbf{x}$ to $\mathbf{y}$. Equation (1) captures radiance changes at fixed support and the spatial mixing that makes boundary evidence uncertain.

3.2 OVERVIEW OF RUBIN

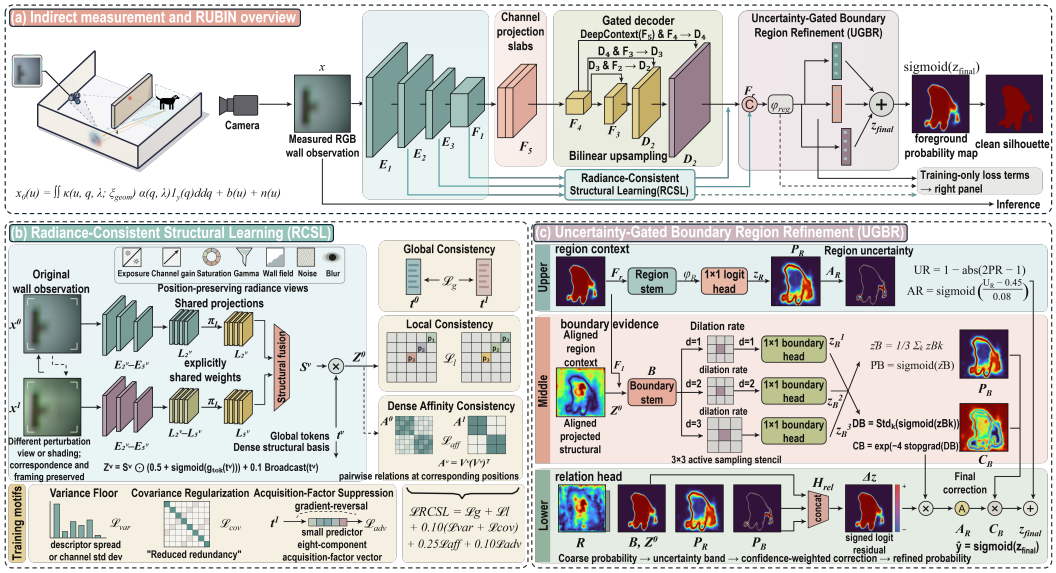

Figure 2: Architecture of RUBIN. (a) An RGB relay wall observation passes through a ResNet50 encoder and gated decoder; structural features from RCSL support UGBR silhouette refinement. (b) RCSL aligns global tokens, co-located features, and dense affinities across position-preserving radiance views, with redundancy reduction and acquisition-factor suppression. (c) UGBR combines coarse regions and multiscale boundary evidence to predict a confidence-weighted logit correction emphasizing uncertain regions. The perturbed view and auxiliary loss branches are used only during training.

Figure 2(a) summarizes RUBIN, the Radiance-Robust Uncertainty-Guided Boundary Inference Network. A ResNet50 encoder produces raw multiscale features $\{\mathbf{E}_l\}_{l=1}^{5}$; channel projections convert

them to decoder features $\{\mathbf{F}_l\}_{l=1}^{5}$, and a gated decoder produces $\mathbf{D}_2$. RCSL extracts a dense structural representation $\mathbf{Z}^0$ from the raw encoder pyramid and fuses it with the decoder features:

$$\mathbf{F}_r = \phi_{\mathrm{reg}}\big([\mathbf{D}_2, \mathbf{Z}^0]\big)\,. \tag{3}$$

Uncertainty-Gated Boundary Region Refinement (UGBR) first predicts coarse regions and multiscale boundaries, then concentrates their interaction at uncertain locations. The exact decoder recurrence is provided in Appendix B.3.

3.3 Radiance-Consistent Structural Learning

Position-preserving radiance views. As illustrated in Figure 2(b), RCSL preserves wall-position correspondence by perturbing radiance on a fixed wall-coordinate grid. Let $\mathbf{I}$ denote the denormalized RGB observation. We model exposure e, channel gain $\mathbf{g}$, saturation s_{sat}, gamma γ, a low-frequency wall field $\rho\mathbf{W}_{\mathrm{wall}}$, noise $\eta\boldsymbol{\epsilon}$, and blur β as follows:

$$\mathbf{I}_{\mathrm{sat}} = \mu_c(\mathbf{I}) + s_{\mathrm{sat}}\big(\mathbf{I} - \mu_c(\mathbf{I})\big), \quad \mathbf{J} = \Big(\mathrm{clip}_{[0,1]}(e\,\mathbf{g} \odot \mathbf{I}_{\mathrm{sat}})\Big)^{\gamma} \odot (1 + \rho\mathbf{W}_{\mathrm{wall}})\,, \tag{4}$$

$$\widetilde{\mathbf{x}} = \mathrm{Norm}_{\mathrm{IM}}\Big(\mathrm{clip}_{[0,1]}((1-\beta)(\mathbf{J} + \eta\boldsymbol{\epsilon}) + \beta\mathcal{A}_3(\mathbf{J} + \eta\boldsymbol{\epsilon}))\Big)\,, \tag{5}$$

Here, μ_c is the channel mean, $\mathcal{A}_3$ is average pooling, and $\mathrm{Norm}_{\mathrm{IM}}$ is ImageNet normalization. These views preserve spatial coordinates while modeling the acquisition variations used during training. All sampling ranges are provided in Appendix B.4.

Global, co-located local, and dense structural representations. The original and perturbed views share the same encoder. For $v \in \{0, 1\}$, we project features from four semantic levels, fuse the aligned feature maps, and retain a global token:

$$\mathbf{L}_l^v = \pi_l(\mathbf{E}_l^v),\ l \in \{2,3,4,5\}, \quad \mathbf{t}^v = h(\mathrm{GAP}(\mathbf{L}_5^v))\,, \tag{6}$$

$$\mathbf{S}^v = \phi_{\mathrm{str}}([\mathbf{L}_2^v, \uparrow\mathbf{L}_3^v, \uparrow\mathbf{L}_4^v, \uparrow\mathbf{L}_5^v])\,, \tag{7}$$

$$\mathbf{Z}^v = \mathbf{S}^v \odot (0.5 + \sigma(g_{\mathrm{tok}}(\mathbf{t}^v))) + 0.1\,\mathrm{Broadcast}(\mathbf{t}^v). \tag{8}$$

By modulating the dense basis, the global token combines scene context with position-specific evidence.

Consistency and statistical regularization. Let superscripts 0 and 1 denote the two views. RCSL aligns their global tokens, corresponding local feature maps, and dense self-affinity relations:

$$\mathcal{L}_{\mathrm{g}} = \frac{1}{B_{\mathrm{sz}} C_{\mathrm{str}}} \sum_{b=1}^{B_{\mathrm{sz}}} \left\| \bar{\mathbf{t}}_b^0 - \bar{\mathbf{t}}_b^1 \right\|_2^2, \tag{9}$$

$$\mathcal{L}_{\mathrm{l}} = \frac{1}{4} \sum_{l=2}^{5} \mathrm{SmoothL1}\big(\mathrm{norm}_c(\mathbf{L}_l^0), \mathrm{norm}_c(\mathbf{L}_l^1)\big)\,, \tag{10}$$

$$\mathcal{L}_{\mathrm{aff}} = \frac{1}{B_{\mathrm{sz}} N_{\mathrm{aff}}^2} \sum_{b=1}^{B_{\mathrm{sz}}} \left\| \mathbf{V}_{\mathrm{aff},b}^0 (\mathbf{V}_{\mathrm{aff},b}^0)^{\top} - \mathbf{V}_{\mathrm{aff},b}^1 (\mathbf{V}_{\mathrm{aff},b}^1)^{\top} \right\|_F^2. \tag{11}$$

Here, B_{sz} and $C_{\mathrm{str}} = 64$ denote the batch size and structural feature dimension. The normalized token is $\bar{\mathbf{t}}_b^v = \mathbf{t}_b^v / \max(\|\mathbf{t}_b^v\|_2, \varepsilon_{\mathrm{g}})$ with $\varepsilon_{\mathrm{g}} = 10^{-12}$. Each $\mathbf{V}_{\mathrm{aff},b}^v$ contains N_{aff} normalized descriptors pooled from $\mathbf{Z}_b^v$, and $\mathrm{SmoothL1}$ averages its elementwise penalty. Variance and covariance penalties encourage descriptor dispersion and reduce channel redundancy, while a predictor with gradient reversal discourages encoding of sampled acquisition factors. Their definitions are provided in Appendix B.6. The complete objective is

$$\mathcal{L}_{\mathrm{RCSL}} = \mathcal{L}_{\mathrm{g}} + \mathcal{L}_{\mathrm{l}} + 0.10(\mathcal{L}_{\mathrm{var}} + \mathcal{L}_{\mathrm{cov}}) + 0.25\mathcal{L}_{\mathrm{aff}} + 0.10\mathcal{L}_{\mathrm{adv}}. \tag{12}$$

3.4 Uncertainty-Gated Boundary Region Refinement

Coarse regions and multiscale boundary evidence. UGBR (Figure 2(c)) first predicts coarse regions, then constructs boundary features from shallow details, region context, and $\mathbf{Z}^0$. Three

branches with dilation rates $\{1,2,3\}$ produce boundary logits $\{\mathbf{z}_B^k\}_{k=1}^3$:

$$\mathbf{R} = \phi_R(\mathbf{F}_r), \quad \mathbf{z}_R = \uparrow w_R(\mathbf{R}), \quad \mathbf{P}_R = \sigma(\mathbf{z}_R), \tag{13}$$

$$\mathbf{B} = \phi_B\big([\pi_d(\mathbf{F}_1), \uparrow\mathbf{R}, \uparrow\pi_{\text{aux}}(\mathbf{Z}^0)]\big)\,,\ \mathbf{H}_B^k = \mathcal{B}_k(\mathbf{B}; k),\ \mathbf{z}_B^k = \uparrow_{(H,W)} w_B^k(\mathbf{H}_B^k)\ (k \in \{1,2,3\}),$$

$$\overline{\mathbf{z}}_B = \tfrac{1}{3}\sum_{k=1}^3 \mathbf{z}_B^k, \quad \mathbf{P}_B = \sigma(\overline{\mathbf{z}}_B), \tag{14}$$

$$\mathbf{D}_B = \sqrt{\tfrac{1}{3}\sum_{k=1}^3 \left(\sigma(\mathbf{z}_B^k) - \tfrac{1}{3}\sum_{j=1}^3 \sigma(\mathbf{z}_B^j)\right)^2}, \quad \mathbf{C}_B = \exp(-4\,\text{stop}(\mathbf{D}_B))\,. \tag{15}$$

Disagreement across boundary scales is summarized by $\mathbf{D}_B$ and converted into boundary confidence $\mathbf{C}_B$.

Relational correction within uncertain regions. Coarse-prediction uncertainty defines a continuous band for correcting ambiguous foreground assignments. A relation head predicts a residual $\Delta\mathbf{z}$ from aligned region, boundary, structural, and probability features:

$$\mathbf{U}_R = 1 - |2\mathbf{P}_R - 1|\,, \quad \mathbf{A}_R = \sigma\left(\frac{\mathbf{U}_R - \tau_{\text{band}}}{T_{\text{band}}}\right), \tag{16}$$

$$\Delta\mathbf{z} = w_{\text{rel}}(\mathbf{H}_{\text{rel}}), \quad \mathbf{z}_{\text{final}} = \mathbf{z}_R + \mathbf{A}_R \odot \mathbf{C}_B \odot \Delta\mathbf{z}. \tag{17}$$

We set $\tau_{\text{band}} = 0.45$ and $T_{\text{band}} = 0.08$. Equation (17) gives greater correction weight to uncertain foreground assignments supported by agreement across boundary scales. The relation inputs are defined in Appendix B.7.

3.5 Training Objective and Inference

We jointly supervise foreground membership, boundary placement, and their relations:

$$\begin{aligned}\mathcal{L}_R &= \text{BCE}(\mathbf{z}_R, \mathbf{y}) + \text{Dice}(\mathbf{P}_R, \mathbf{y}),\\ \mathbf{y}_B &= \text{Edge}_3(\mathbf{y}), \quad \mathcal{L}_B = \text{wBCE}(\overline{\mathbf{z}}_B, \mathbf{y}_B) + \text{Dice}(\mathbf{P}_B, \mathbf{y}_B),\end{aligned} \tag{18}$$

$$\begin{aligned}\mathcal{L}_{\text{rel}} = \frac{1}{B_{\text{sz}}HW}\,&\|\mathbf{P}_B - \text{Edge}_3(\sigma(\mathbf{z}_{\text{final}}))\|_1\\ &+ \frac{0.5}{B_{\text{sz}}HW}\,\|\text{Edge}_3(\sigma(\mathbf{z}_{\text{final}})) - \mathbf{y}_B\|_1\,.\end{aligned} \tag{19}$$

Here, Edge_3 denotes the 3×3 morphological gradient, and wBCE reweights sparse positive boundary pixels. We match the scale-normalized soft Euler surrogate $\mathfrak{E}$ on a grid of at most 64×64 to provide a weak constraint against fragmentation and spurious holes:

$$\mathcal{L}_{\text{top}} = \text{SmoothL1}(\mathfrak{E}(\sigma(\mathbf{z}_{\text{final}})), \mathfrak{E}(\mathbf{y}))\,. \tag{20}$$

This term provides auxiliary supervision on a scale-normalized component-and-hole statistic. The auxiliary losses are detailed in Appendix B.8.

Let $\mathcal{S}$ be the indexed collection of the final logits, coarse logits, and logits from three decoder side outputs, with $\mathcal{S}_k$ denoting entry k. The complete objective is

$$\mathcal{L}_{\text{UGBR}} = 0.50\mathcal{L}_R + 0.75\mathcal{L}_B + 0.30\mathcal{L}_{\text{rel}} + 0.02\mathcal{L}_{\text{top}}, \tag{21}$$

$$\mathcal{L}_{\text{seg}} = \sum_{k=1}^{|\mathcal{S}|}\left[\text{BCE}(\mathcal{S}_k, \mathbf{y}) + \mathcal{L}_{\text{IoU}}(\sigma(\mathcal{S}_k), \mathbf{y})\right], \tag{22}$$

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + 0.05\mathcal{L}_{\text{RCSL}} + 0.25\mathcal{L}_{\text{UGBR}}. \tag{23}$$

At inference, RUBIN maps one normalized RGB relay wall observation to the target foreground probability map $\sigma(\mathbf{z}_{\text{final}})$; radiance-view generation and auxiliary objectives serve training. Target masks supervise training and provide the reference for evaluating region overlap and pixel error.

4 Experiments

We first test which aspects of hidden-target shape indirect wall light can reveal. Ablations then examine the contributions of structural learning and boundary refinement, before observation-variation experiments assess the shape information retained as wall-light patterns change.

Table 1: Quantitative comparison on five measured NLOS benchmarks using Max-F↑, MAE↓, and Dice↑. The best results are highlighted in **red**, and the second-best results in **blue**. Tied values share the same rank. Blue row bands highlight GateNet and ZoomNeXt, which rank second and third overall according to their average rank across all fifteen metrics.

Method	Venue	NLOS-HKU			NLOS-ECSSD			NLOS-DUTS			NLOS-DUT-OMRON			NLOS-PASCAL		
		Max-F↑	MAE↓	Dice↑	Max-F↑	MAE↓	Dice↑	Max-F↑	MAE↓	Dice↑	Max-F↑	MAE↓	Dice↑	Max-F↑	MAE↓	Dice↑
Direct-view segmentation																
LDF	CVPR'20	**0.527**	0.178	**0.463**	0.604	0.194	0.517	0.651	0.151	0.612	0.420	0.147	0.315	0.412	0.235	0.338
ZoomNeXt	**TPAMI'24**	0.520	0.178	0.434	0.614	0.199	0.557	0.674	0.143	0.630	0.422	0.148	0.333	0.451	0.246	0.389
GateNet	**ECCV'20**	0.519	0.176	0.442	0.621	0.184	**0.565**	0.673	0.145	**0.639**	**0.434**	0.145	**0.340**	0.417	0.229	0.343
FAP-Net	TIP'22	0.512	0.182	0.413	0.605	0.191	0.525	0.671	0.146	0.627	0.393	0.146	0.261	0.415	0.234	0.331
PGNet	CVPR'22	0.499	**0.172**	0.391	**0.625**	**0.175**	0.531	0.619	0.160	0.573	0.188	0.151	0.000	0.463	0.214	0.401
BiRefNet	CAAI AIR'24	0.508	0.173	0.411	0.586	0.199	0.497	0.632	0.150	0.586	0.188	0.151	0.000	0.458	0.216	0.387
SegFormer	NeurIPS'21	0.466	0.192	0.350	0.561	0.209	0.452	0.571	0.154	0.523	0.188	0.144	0.115	0.458	**0.210**	0.391
CamoFormer	TPAMI'24	0.489	0.193	0.346	0.581	0.200	0.476	0.614	0.178	0.591	0.378	0.160	0.209	0.469	0.213	0.378
VSCode	CVPR'24	0.473	0.198	0.423	0.571	0.205	0.514	0.614	0.162	0.559	0.399	0.154	0.301	0.451	0.221	0.384
Indirect-imaging methods																
NLOS-OT	TIP'22	0.509	0.175	0.432	0.601	0.184	0.541	0.670	**0.137**	0.618	0.423	**0.137**	0.328	0.404	0.242	0.352
Learnable Inv. Kernel	ICCV'23	0.505	0.176	0.428	0.602	0.178	0.524	0.660	0.142	0.610	0.428	0.139	0.335	0.424	0.229	0.365
NeTF	TPAMI'21	0.511	**0.172**	0.425	0.600	0.182	0.515	0.673	**0.137**	0.615	0.415	0.140	0.322	0.418	0.233	0.367
NLOS-LTM	TIP'25	0.490	0.175	0.369	0.607	0.182	0.502	0.635	0.156	0.598	0.188	0.151	0.000	0.444	0.237	0.385
PAC-Net	CVPR'23	0.469	0.186	0.357	0.584	0.188	0.483	0.614	0.164	0.588	0.277	0.148	0.138	**0.476**	0.216	**0.409**
NLOST	CVPR'23	0.522	**0.172**	0.447	0.590	0.185	0.533	**0.676**	**0.137**	0.621	0.387	0.139	0.284	0.394	0.226	0.308
LEAP	ECCV'24	0.496	0.191	0.417	0.602	0.191	0.534	0.621	0.173	0.589	0.403	0.160	0.260	0.448	0.224	0.373
NLOS-LPP	ICCV'25	0.498	0.192	0.417	0.590	0.191	0.514	0.626	0.170	0.591	0.402	0.161	0.266	0.475	0.224	0.408
FMoE-NLOS	PR'26	0.492	0.191	0.419	0.600	0.191	0.522	0.620	0.171	0.583	0.405	0.160	0.269	0.474	0.215	0.408
ST-Mamba	NeurIPS'24	0.500	0.192	0.420	0.601	0.190	0.523	0.625	0.170	0.597	0.410	0.161	0.270	0.447	0.239	0.365
TransiT	ICCV'25	0.489	0.191	0.414	0.593	0.189	0.518	0.624	0.174	0.594	0.388	0.161	0.258	0.450	0.233	0.381
Ours																
RUBIN	**Ours**	**0.565**	**0.161**	**0.497**	**0.751**	**0.109**	**0.697**	**0.692**	**0.136**	**0.640**	**0.447**	**0.134**	**0.361**	**0.488**	**0.209**	**0.410**

4.1 Experimental Setup

Datasets and acquisition protocol. We evaluate RUBIN on five measured NLOS subsets derived from standard salient-object collections (Li & Yu, 2015; Yan et al., 2013; Wang et al., 2017; Yang et al., 2013; Li et al., 2014): NLOS-HKU, NLOS-ECSSD, NLOS-DUTS, NLOS-DUT-OMRON, and NLOS-PASCAL. The comparison table includes direct-view baselines and NLOS architectures from prior work (Wei et al., 2020; Pang et al., 2024; Zhao et al., 2020; Zhou et al., 2022; Xie et al., 2022; Zheng et al., 2024; Xie et al., 2021; Yin et al., 2024; Luo et al., 2024; Geng et al., 2022; Yu et al., 2023; Shen et al., 2021; Zhang et al., 2025; Wang et al., 2023; Li et al., 2023a; Cho et al., 2024; Sun et al., 2025; Zhang et al., 2026; Li et al., 2024a; 2025). During data acquisition, a planar target source positioned around a corner and occluded by a T-shaped baffle produces two-dimensional target patterns. The emitted light forms indirect patterns on a visible relay wall. A Sony α7S II camera equipped with a TTArtisan 50 mm f/1.4 ASPH lens observes a matte white relay wall and captures the RGB light pattern corresponding to each target. Foreground annotations of the target images serve as the target ground-truth masks. Camera settings and scene geometry remain fixed in the reference configuration. The reported training/test counts of paired wall observations and target masks are 4,007/440, 1,000/200, 9,497/1,055, 4,652/516, and 765/85, respectively. Models, augmented views, and representation probes use only relay wall observations; paired target masks provide supervision and evaluation on fixed acquisition partitions.

Photographs of the physical acquisition scene, alignment references, and camera support are provided in Figure 6 and described in Appendix A.

Baselines and comparison protocol. We compare RUBIN with twenty representative methods covering direct-view segmentation and indirect imaging (Table 1). Each comparison architecture is evaluated on predicting a target mask from an RGB wall observation. Where the original sensing or prediction interface differs, the input and output interfaces are adapted while retaining the principal representation modules. All methods use the same data partitions, input resolution, target-mask supervision, and evaluation procedure. These results compare the adapted architectures within the RGB-to-mask task.

Implementation and evaluation. RUBIN is implemented in PyTorch with a ResNet50 encoder. Training images and masks are resized to 384×384. The base decoder width and structural dimension are both 64. We train for 100 epochs using Adam on a single NVIDIA GeForce RTX 3080, with a batch size of 5 and a random seed of 3407. The initial learning rate is 10^{-3}, with 10^{-4} for the encoder, and the weight decay is 5×10^{-4}. Training uses the objective in Eq. (23), and we eval-

uate the checkpoint from epoch 100. We report the maximum F-measure over thresholds (Max-F), mean absolute error (MAE), and Dice coefficient. For evaluation, each foreground probability map is bilinearly resized to the ground-truth mask resolution, normalized, and quantized to 8 bits before metric accumulation. Higher Max-F and Dice and lower MAE indicate better performance. Metric descriptions are given in Appendix D.1.

4.2 Comparison with Representative Methods

RUBIN achieves the highest Max-F, the lowest MAE, and the highest Dice on all five datasets in Table 1. The clearest separation appears on NLOS-ECSSD, where Max-F increases from 0.625 to 0.751 and MAE decreases from 0.175 to 0.109 relative to the strongest competing result in each metric. The joint improvement indicates better foreground recovery together with lower pixelwise error.

RUBIN improves foreground overlap on all five collections, with dataset-dependent gains. Relative to the strongest competing Dice score on each dataset, RUBIN gains 0.034 on NLOS-HKU, 0.132 on NLOS-ECSSD, 0.001 on NLOS-DUTS, 0.021 on NLOS-DUT-OMRON, and 0.001 on NLOS-PASCAL. These gains indicate closer foreground agreement; margins are narrow on NLOS-DUTS and NLOS-PASCAL. Max-F evaluates the best thresholded foreground separation, MAE measures probability-map error, and Dice measures foreground overlap under the stated evaluation procedure. Together, these measures connect stronger foreground separation with greater overlap and lower pixelwise error. The corresponding masks show what these numerical gains mean for the shape behind the corner.

GateNet regulates cross-level feature transmission (Zhao et al., 2020), while ZoomNeXt combines mixed-scale features with an uncertainty-aware training loss (Pang et al., 2024). RUBIN couples radiance-consistent structure with uncertainty-gated boundary and region correction, achieving a Dice score of 0.697 on NLOS-ECSSD compared with 0.565 for GateNet and 0.557 for ZoomNeXt in this wall-observation comparison. The five-dataset results extend the advantage across different target collections. Appendix D examines the recovered structure with additional metrics and visual comparisons.

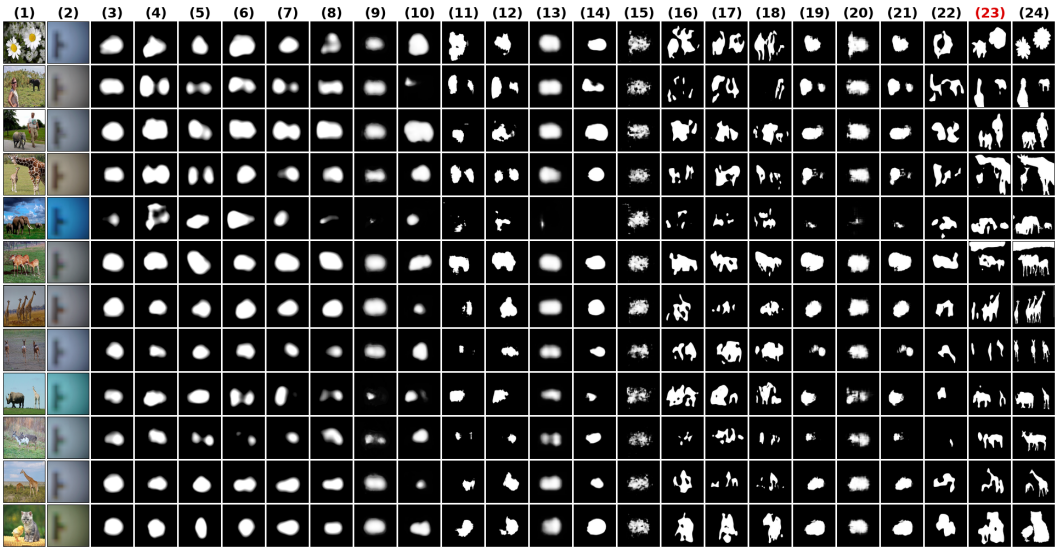

Figure 3: Qualitative comparison on measured NLOS observations. From left to right, the columns show: (1) target source image, (2) relay wall projection, (3) LDF, (4) ZoomNeXt, (5) GateNet, (6) FAP-Net, (7) PGNet, (8) BiRefNet, (9) SegFormer, (10) CamoFormer, (11) VSCode, (12) NLOS-OT, (13) Learnable Inverse Kernel, (14) NeTF, (15) NLOS-LTM, (16) PAC-Net, (17) NLOST, (18) LEAP, (19) NLOS-LPP, (20) FMoE-NLOS, (21) ST-Mamba, (22) TransiT, (23) RUBIN (ours), and (24) target ground-truth mask. Column 1 provides the target reference, and column 2 supplies the wall observation used by every evaluated model.

RUBIN recovers target-specific shape from mixed wall-light patterns (Figure 3). Several comparison methods produce similar rounded responses for targets with different structures, while others

divide the foreground into irregular fragments. RUBIN more often recovers the target's main extent, dominant orientation, and separation between major components. These observations help interpret the overlap gains in Table 1: the predictions retain more of the hidden target's spatial structure and provide a useful outline for preliminary inference around corners.

4.3 Ablation Studies

Table 2: Ablation study on NLOS-ECSSD. The best results are highlighted in **red**, and the second-best results in **blue**. Tied values share the same rank. Full RUBIN serves as the full-model reference matching Table 1. Higher Max-F and Dice and lower MAE indicate better performance.

Variant	Max-F↑	MAE↓	Dice↑
Component contributions			
Without RCSL and UGBR	0.594	0.189	0.528
Without UGBR	0.674	0.149	0.617
Without RCSL	0.696	0.138	0.637
Full RUBIN	**0.751**	**0.109**	**0.697**
RCSL mechanisms			
Without global consistency	0.728	**0.124**	**0.674**
Generic photometric views	**0.733**	0.125	0.658
UGBR variants and refiner control			
Generic refiner replacing UGBR	0.701	0.136	0.637
Relation ablation	0.727	**0.124**	0.668
Coarse-region ablation	0.709	0.133	0.643

Contributions of RCSL and UGBR. RCSL and UGBR improve successive stages: learning stable structure and forming coherent foreground regions. Starting from a baseline Dice of 0.528, the RCSL-only and UGBR-only variants reach 0.617 and 0.637, respectively; combining them raises Dice to 0.697 (Table 2). Relative to the stronger single-component variant, the full model improves Max-F by 0.055 and Dice by 0.060, while reducing MAE by 0.029. RCSL constrains decoder features, while UGBR corrects coarse foreground assignments. Their joint gain supports carrying radiance-consistent structure through to uncertainty-guided foreground correction; the component comparison evaluates the combined designs rather than isolating every auxiliary loss.

RCSL mechanisms. Global consistency raises Max-F from 0.728 to 0.751, while the proposed radiance views raise Dice from 0.658 to 0.697 relative to generic photometric augmentation. These changes connect foreground recovery to both cross-view structural agreement and the modeling of acquisition-related radiance during training.

UGBR mechanisms and refiner control. UGBR raises Dice from the generic convolutional refiner's 0.637 to 0.697. The reported relation and coarse-region ablations yield Dice scores of 0.668 and 0.643, respectively. In the ablation protocol, disabling the relation switch removes both residual correction and its auxiliary loss. Disabling the coarse-region switch instead derives the region seed from boundary features and removes the UGBR region auxiliary loss, while retaining region context and a coarse output entry. Each switch jointly changes the corresponding prediction path and auxiliary supervision. The comparison supports UGBR's joint use of coarse regions, uncertainty, and multiscale boundary evidence; matched capacity is not established.

4.4 Robustness to Acquisition Variations

Evaluation protocol. The qualitative comparison in Figure 5 covers the reported pitch, yaw, and roll range of $[-12°, +12°]$, camera distances from 2.5 m to 7.0 m with a 4.5 m reference, source brightness from 300 to 600 nits, and rough textured, diffuse, brick, and glossy relay surfaces. Figure 4 illustrates the corresponding acquisition layout. Appendix E separately evaluates computational perturbations of recorded wall observations under fixed model parameters.

Qualitative robustness. We compare RUBIN, GateNet, ZoomNeXt, and PAC-Net under representative observation variations (Figure 5). RUBIN more often retains the dominant elongated foreground and its main extent, so its predictions continue to convey the target's orientation and occupied

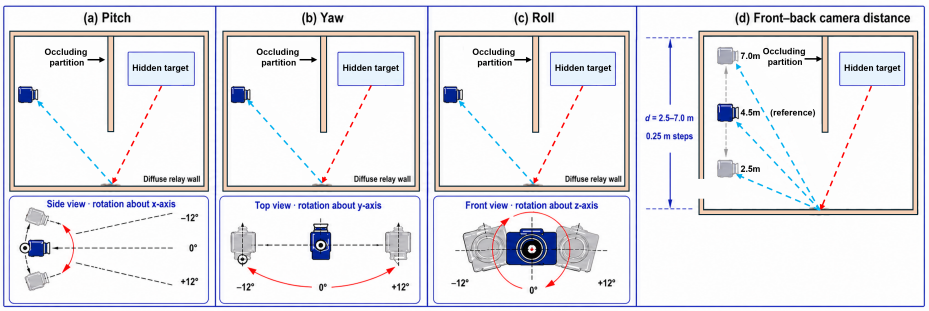

Figure 4: Experimental layout for the qualitative robustness study. Camera pitch, yaw, and roll are varied independently from $-12°$ to $+12°$; camera distance is varied from 2.5 m to 7.0 m in 0.25 m increments around a 4.5 m reference.

region as the wall-light pattern changes. Together with the component ablations, these observations are consistent with the roles of radiance-consistent features and uncertainty-guided refinement. Appendix E provides all-setting visualizations and four-metric summaries for computational geometry, radiance, and JPEG perturbations.

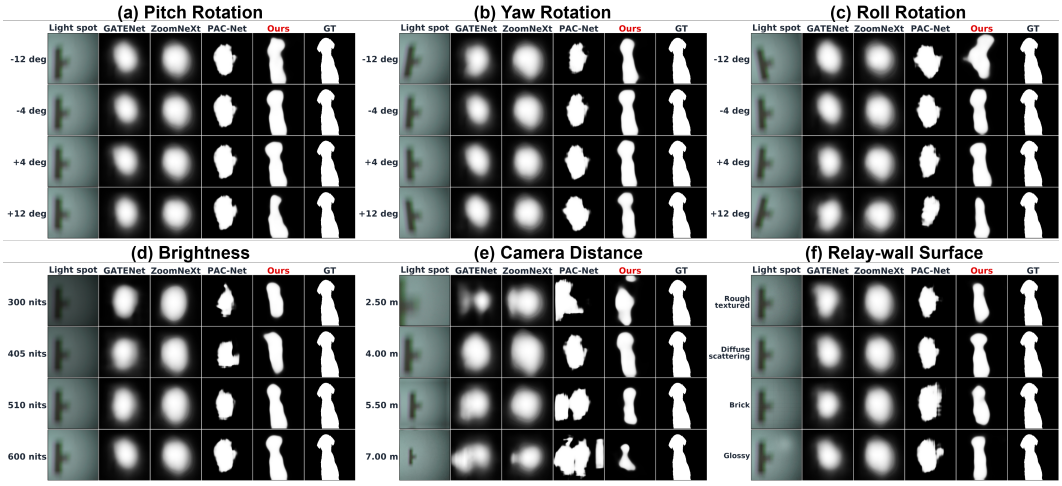

Figure 5: Qualitative robustness comparison under acquisition variations. The upper groups show results under pitch, yaw, and roll rotations. The lower groups evaluate changes in source brightness, camera distance, and relay wall material. Each group shows the wall-light pattern, predictions from GateNet, ZoomNeXt, PAC-Net, and RUBIN, and the target ground-truth mask.

Shape information across configurations. Across the configurations, RUBIN generally retains more of the target's main extent and orientation than the comparison methods. Large distance changes, rotations, and glossy relay responses can reduce foreground extent or distort boundaries. The comparison supports preliminary recovery of basic target shape as the observed wall-light pattern changes.

5 Conclusion

RUBIN demonstrates the feasibility of recovering basic hidden-target shape from indirect wall light with a conventional camera. RCSL learns structure under modeled radiance changes; UGBR refines foreground regions using uncertainty and multiscale boundary agreement. Experiments on five measured datasets show an overall advantage over twenty representative methods, with ablations supporting the components' complementary roles. For the evaluated planar target patterns, foreground extent, contours, and component arrangement support preliminary shape inference around corners. Future work will extend this capability across broader scene geometries, relay responses, and hidden-target configurations.

REFERENCES

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICRegL: Self-supervised learning of local visual features. In *Advances in Neural Information Processing Systems*, 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/39cee562b91611c16ac0b100f0bc1ea1-Abstract-Conference.html.

Katherine L. Bouman, Vickie Ye, Adam B. Yedidia, Fredo Durand, Gregory W. Wornell, Antonio Torralba, and William T. Freeman. Turning Corners into Cameras: Principles and Methods. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2289–2297, 2017. doi: 10.1109/iccv.2017.249.

In Cho, Hyunbo Shim, and Seon Joo Kim. Learning to Enhance Aperture Phasor Field for Non-Line-of-Sight Imaging. In *European Conference on Computer Vision*, pp. 72–89, 2024. doi: 10.1007/978-3-031-72775-7_5.

Xingyu Cui, Huanjing Yue, Song Li, Xiangjun Yin, Yusen Hou, Yun Meng, Kai Zou, Xiaolong Hu, and Jingyu Yang. Virtual scanning: Unsupervised non-line-of-sight imaging from irregularly undersampled transients. In *Advances in Neural Information Processing Systems*, volume 37, pp. 109381–109406, 2024. doi: 10.52202/079017-3472.

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-Measure: A New Way to Evaluate Foreground Maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4558–4567, 2017. doi: 10.1109/iccv.2017.487.

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 698–704, 2018. doi: 10.24963/ijcai.2018/97.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL https://jmlr.org/papers/v17/15-239.html.

Ruixu Geng, Yang Hu, Zhi Lu, Cong Yu, Houqiang Li, Hengyu Zhang, and Yan Chen. Passive non-line-of-sight imaging using optimal transport. *IEEE Transactions on Image Processing*, 31: 110–124, 2022. doi: 10.1109/TIP.2021.3128312.

Saumya Gupta, Yikai Zhang, Xiaoling Hu, Prateek Prasanna, and Chao Chen. Topology-aware uncertainty for image segmentation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8186–8207, 2023. doi: 10.52202/075280-0359.

Connor Hashemi, Rafael Avelar, and James Leger. Isolating signals in passive non-line-of-sight imaging using spectral content. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7328–7339, 2025. doi: 10.1109/TPAMI.2023.3301336.

Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5455–5463, 2015. doi: 10.1109/cvpr.2015.7299184.

Ruiqian Li, Siyuan Shen, Suan Xia, Ziheng Wang, Xingyue Peng, Chengxuan Song, Yingsheng Zhu, Tao Wu, Shiying Li, and Jingyi Yu. TransiT: Transient Transformer for Non-Line-of-Sight Videography. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27542–27551, 2025. doi: 10.1109/iccv51701.2025.02557.

Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The Secrets of Salient Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 280–287, 2014. doi: 10.1109/cvpr.2014.43.

Yue Li, Jiayong Peng, Juntian Ye, Yueyi Zhang, Feihu Xu, and Zhiwei Xiong. NLOST: Non-Line-of-Sight Imaging with Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13313–13322, 2023a. doi: 10.1109/cvpr52729.2023.01279.

Yue Li, Yueyi Zhang, Juntian Ye, Feihu Xu, and Zhiwei Xiong. Deep non-line-of-sight imaging from under-scanning measurements. In *Advances in Neural Information Processing Systems*, volume 36, pp. 59095–59106, 2023b. doi: 10.52202/075280-2578.

Yue Li, Yi Sun, Shida Sun, Juntian Ye, Yueyi Zhang, Feihu Xu, and Zhiwei Xiong. Toward dynamic non-line-of-sight imaging with mamba enforced temporal consistency. In *Advances in Neural Information Processing Systems*, volume 37, pp. 126452–126473, 2024a. doi: 10.52202/079017-4016.

Yunheng Li, Zhong-Yu Li, Quan-Sheng Zeng, Qibin Hou, and Ming-Ming Cheng. Cascade-CLIP: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28243–28258. PMLR, 2024b. URL `https://proceedings.mlr.press/v235/li24aq.html`.

Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. VSCode: General Visual Salient and Camouflaged Object Detection with 2D Prompt Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17169–17180, 2024. doi: 10.1109/cvpr52733.2024.01625.

Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to Evaluate Foreground Maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2014. doi: 10.1109/cvpr.2014.39.

Fangzhou Mu, Sicheng Mo, Jiayong Peng, Xiaochun Liu, Ji Hyun Nam, Siddeshwar Raghavan, Andreas Velten, and Yin Li. Physics to the rescue: Deep non-line-of-sight reconstruction for high-speed imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8): 6146–6158, 2025. doi: 10.1109/TPAMI.2022.3203383.

Ji Hyun Nam, Eric Brandt, Sebastian Bauer, Xiaochun Liu, Marco Renna, Alberto Tosi, Eftychios Sifakis, and Andreas Velten. Low-latency time-of-flight non-line-of-sight imaging at 5 frames per second. *Nature Communications*, 12, 2021. doi: 10.1038/s41467-021-26721-x.

Matthew O'Toole, David B. Lindell, and Gordon Wetzstein. Confocal non-line-of-sight imaging based on the light-cone transform. *Nature*, 555(7696):338–341, 2018. doi: 10.1038/nature25489.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. ZoomNeXt: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9205–9220, 2024. doi: 10.1109/TPAMI.2024.3417329.

Charles Saunders, John Murray-Bruce, and Vivek K. Goyal. Computational periscopy with an ordinary digital camera. *Nature*, 565(7740):472–475, 2019. doi: 10.1038/s41586-018-0868-6.

Jan Philipp Schneider, Mishal Fatima, Jovita Lukasik, Andreas Kolb, Margret Keuper, and Michael Moeller. Implicit representations for constrained image segmentation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43765–43790. PMLR, 2024. URL `https://proceedings.mlr.press/v235/schneider24a.html`.

Siyuan Shen, Zi Wang, Ping Liu, Zhengqing Pan, Ruiqian Li, Tian Gao, Shiying Li, and Jingyi Yu. Non-line-of-Sight Imaging via Neural Transient Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:2257–2268, 2021. doi: 10.1109/tpami.2021.3076062.

Ravid Shwartz-Ziv, Randall Balestriero, Kenji Kawaguchi, Tim G. J. Rudner, and Yann LeCun. An information theory perspective on variance-invariance-covariance regularization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 33965–33998, 2023. doi: 10.52202/075280-1472.

Shida Sun, Yue Li, Yueyi Zhang, and Zhiwei Xiong. Generalizable Non-Line-of-Sight Imaging with Learnable Physical Priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–10, 2025. doi: 10.1109/iccv51701.2025.02322.

Alex N. Wang, Christopher Hoang, Yuwen Xiong, Yann LeCun, and Mengye Ren. PooDLe: Pooled and dense self-supervised learning from naturalistic videos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=dEg5SdGaiq.

Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to Detect Salient Objects with Image-Level Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3796–3805, 2017. doi: 10.1109/cvpr.2017.404.

Yihao Wang, Zhigang Wang, Bin Zhao, Dong Wang, Mulin Chen, and Xuelong Li. Propagate and Calibrate: Real-Time Passive Non-Line-of-Sight Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 972–981, 2023. doi: 10.1109/cvpr52729.2023.00100.

Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label Decoupling Framework for Salient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13022–13031, 2020. doi: 10.1109/cvpr42600.2020.01304.

Chenxi Xie, Changqun Xia, Mingcan Ma, Zhirui Zhao, Xiaowu Chen, and Jia Li. Pyramid Grafting Network for One-Stage High Resolution Saliency Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11707–11716, 2022. doi: 10.1109/cvpr52688.2022.01142.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2105.15203.

Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical Saliency Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1155–1162, 2013. doi: 10.1109/cvpr.2013.153.

Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency Detection via Graph-Based Manifold Ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3166–3173, 2013. doi: 10.1109/cvpr.2013.407.

Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. CamoFormer: Masked Separable Attention for Camouflaged Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:10362–10374, 2024. doi: 10.1109/tpami.2024.3438565.

Yanhua Yu, Siyuan Shen, Zi Wang, Binbin Huang, Yuehan Wang, Xingyue Peng, Suan Xia, Ping Liu, Ruiqian Li, and Shiying Li. Enhancing Non-line-of-sight Imaging via Learnable Inverse Kernel and Attention Mechanisms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10529–10539, 2023. doi: 10.1109/iccv51070.2023.00969.

Jiarui Zhang, Ruixu Geng, Xiaolong Du, Yan Chen, Houqiang Li, and Yang Hu. Passive non-line-of-sight imaging with light transport modulation. *IEEE Transactions on Image Processing*, 34:410–424, 2025. doi: 10.1109/TIP.2024.3518097.

Qinghua Zhang, Xi Ling, Yuping Duan, and Jun Liu. Frequency-domain multi-regularization-experts fusion for robust non-line-of-sight imaging. *Pattern Recognition*, 173:112914, 2026. doi: 10.1016/j.patcog.2025.112914.

Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and Balance: A Simple Gated Network for Salient Object Detection. In *European Conference on Computer Vision*, pp. 35–51, 2020. doi: 10.1007/978-3-030-58536-5_3.

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral Reference for High-Resolution Dichotomous Image Segmentation. *CAAI Artificial Intelligence Research*, pp. 9150038, 2024. doi: 10.26599/air.2024.9150038.

Tao Zhou, Yi Zhou, Chen Gong, Jian Yang, and Yu Zhang. Feature Aggregation and Propagation Network for Camouflaged Object Detection. *IEEE Transactions on Image Processing*, 31:7036–7047, 2022. doi: 10.1109/tip.2022.3217695.

A Physical Acquisition Scene

Figure 6 shows the physical arrangement used to record indirect wall light.

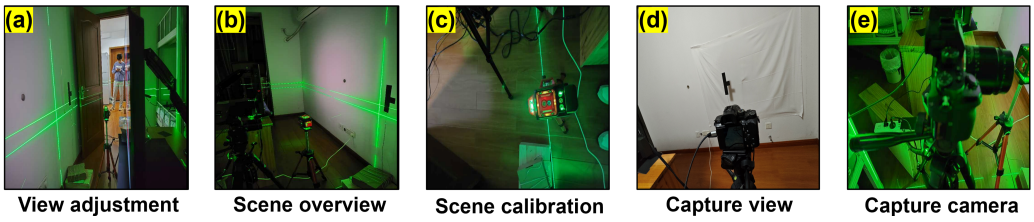

Figure 6: Photographs of the physical acquisition scene and apparatus. (a) Doorway and wall arrangement at the occluding corner. (b) Camera and relay-wall arrangement with visible laser reference lines. (c) Laser level positioned on the floor. (d) Camera facing the white relay surface. (e) Close-up of the camera support and orientation-adjustment hardware.

The photographs show the occluding corner, the camera facing the relay wall, the laser-level references, and the camera support. The reference lines assist scene alignment, while the support allows adjustment of the camera orientation. These geometric relationships determine how indirect light is observed on the wall. Figure 4 illustrates the acquisition-variation layout, and Appendix E examines computational perturbations of the recorded observations.

Acquisition-variation layout. Figure 4 illustrates the camera axes and distance range used in the main qualitative robustness comparison. Each factor is varied independently around the reference configuration. The full image-space perturbation protocol for the extended quantitative sweeps is specified in Appendix E.1.

B Mathematical Details of RUBIN

This appendix details the representations, operators, and objectives used by RUBIN. Bold lowercase letters denote vectors or single-channel maps, bold uppercase letters denote multichannel tensors or matrices, calligraphic letters denote objectives or sets, and plain letters denote scalars or indices. Superscripts 0 and 1 identify the original and radiance-perturbed views.

B.1 Notation, dimensions, and operators

Let B_{sz} be the mini-batch size, let (H, W) be the output resolution, and let $C_{\mathrm{str}} = 64$ be the number of structural channels. The base decoder width is also 64. The paired training tensors are

$$\mathbf{x} \in \mathbb{R}^{B_{\mathrm{sz}} \times 3 \times H \times W}, \qquad \mathbf{y} \in \{0,1\}^{B_{\mathrm{sz}} \times 1 \times H \times W}. \tag{24}$$

Equation (24) states that each training example contains one normalized RGB relay-wall observation and one binary target mask. The singleton channel of $\mathbf{y}$ distinguishes a mask from a three-channel observation.

For a scalar u_{op}, the logistic map, clipping operator, and stop-gradient operator are defined by

$$\begin{aligned} \sigma(u_{\mathrm{op}}) &= \frac{1}{1+\exp(-u_{\mathrm{op}})}, \\ \mathrm{clip}_{[0,1]}(u_{\mathrm{op}}) &= \min\{1, \max\{0, u_{\mathrm{op}}\}\}, \\ \mathrm{stop}(u_{\mathrm{op}}) &= u_{\mathrm{op}} \quad \text{in the forward pass}, \\ \frac{\partial\, \mathrm{stop}(u_{\mathrm{op}})}{\partial u_{\mathrm{op}}} &= 0 \quad \text{in the backward pass}. \end{aligned} \tag{25}$$

Equation (25) collects three basic operators. The first converts a logit into a probability, the second clips values to $[0, 1]$, and the third preserves its input value in the forward pass while blocking its

gradient in the backward pass. All three operators are applied elementwise when their arguments are tensors.

We use $[\cdot,\cdot]$ for channel-wise concatenation, $\odot$ for elementwise multiplication, $\uparrow_{(H_{\mathrm{dst}},W_{\mathrm{dst}})}$ for bilinear resizing to $(H_{\mathrm{dst}}, W_{\mathrm{dst}})$, GAP for global average pooling, and $\|\cdot\|_F$ for the Frobenius norm. The notation $\|\cdot\|_1$ denotes the sum of absolute values over all tensor entries. The encoder-level index is $l \in \{1,2,3,4,5\}$, the view index is $v \in \{0,1\}$, and the boundary-branch index is $k \in \{1,2,3\}$.

B.2 From indirect transport to mask prediction

The target mask induces the following foreground-support indicator on the hidden target plane:

$$\mathbf{1}_{\mathbf{y}}(\mathbf{q}) = \begin{cases} 1, & \text{if the target-plane coordinate } \mathbf{q} \text{ belongs to the foreground,} \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

Equation (26) selects the foreground contribution to the target-plane image. The field $\mathbf{a}(\mathbf{q},\lambda)$ gives its spectral radiance, while the background contribution is included in $\mathbf{b}$.

Under linear superposition of steady-state light transport, the raw RGB response at a visible-wall coordinate $\mathbf{u}$ is

$$\mathbf{x}_0(\mathbf{u}) = \int_{\Omega_{\mathrm{tar}}} \int_{\Lambda} \kappa(\mathbf{u},\mathbf{q},\lambda;\boldsymbol{\xi}_{\mathrm{geom}})\mathbf{a}(\mathbf{q},\lambda)\mathbf{1}_{\mathbf{y}}(\mathbf{q})\,\mathrm{d}\lambda\,\mathrm{d}\mathbf{q} + \mathbf{b}(\mathbf{u}) + \mathbf{n}(\mathbf{u}). \tag{27}$$

Equation (27) is Eq. (1) written as a generative measurement model. The kernel κ aggregates hidden-to-wall propagation, wall reflection, wall-to-camera propagation, and sensor response under the fixed geometry $\boldsymbol{\xi}_{\mathrm{geom}}$. The field $\mathbf{b}$ contains the relay response from the background and ambient illumination, and $\mathbf{n}$ represents measurement noise. The integral shows how foreground support, appearance, and transport jointly shape the observation: many combinations can produce similar wall measurements.

Let m_c and d_c denote the fixed ImageNet mean and standard deviation for RGB channel c. After scaling the stored image to $[0,1]$, the network input is

$$\begin{aligned} x_{b,c,i,j} &= \frac{\mathrm{clip}_{[0,1]}(x_{0,b,c,i,j}) - m_c}{d_c}, \\ (m_1,m_2,m_3) &= (0.485, 0.456, 0.406), \\ (d_1,d_2,d_3) &= (0.229, 0.224, 0.225). \end{aligned} \tag{28}$$

Equation (28) defines the normalized tensor $\mathbf{x}$ used by the encoder. The constants m_c and d_c are the fixed channel-wise normalization statistics.

The network produces a full-resolution logit map and then applies the logistic map:

$$\mathbf{z}_{\mathrm{final}} = f_\theta(\mathbf{x}), \qquad \widehat{\mathbf{y}} = \sigma(\mathbf{z}_{\mathrm{final}}). \tag{29}$$

Equation (29) expands Eq. (2). The parameter set θ contains the encoder, decoder, RCSL, and UGBR parameters, while $\widehat{\mathbf{y}}$ contains foreground probabilities in $[0,1]$.

B.3 Encoder and gated decoder recurrence

The ResNet50 encoder returns five pyramid tensors $\{\mathbf{E}_l\}_{l=1}^{5}$. Channel-reduction operators Red_l and the deep context operator $\mathcal{C}_{\mathrm{ctx}}$ produce

$$\mathbf{F}_l = \mathrm{Red}_l(\mathbf{E}_l), \quad l \in \{1,2,3,4\}, \qquad \mathbf{F}_5 = \mathcal{C}_{\mathrm{ctx}}(\mathrm{Red}_5(\mathbf{E}_5))\,. \tag{30}$$

Equation (30) defines the five decoder-ready feature levels. Each Red_l is the corresponding channel-reduction operator. $\mathcal{C}_{\mathrm{ctx}}$ is the composition of residual convolution blocks with dilation factors two and three, so the deepest feature receives enlarged contextual support before upsampling.

For decoder level $l \in \{4, 3, 2\}$, first resize the deeper tensor and infer a one-channel skip gate:

$$\widetilde{\mathbf{D}}_{l+1} = \uparrow_{(H_l, W_l)} \mathbf{D}_{l+1}, \qquad \mathbf{M}_l = \sigma\Big(w_{l,2} * \mathrm{GELU}\Big(\mathrm{GN}\Big(w_{l,1} * [\widetilde{\mathbf{D}}_{l+1}, \mathbf{F}_l]\Big)\Big)\Big). \tag{31}$$

Equation (31) makes the spatial gate inside each upsampling block explicit. Here, $w_{l,1}$ and $w_{l,2}$ are the first and second gate convolutions, $*$ denotes convolution, and (H_l, W_l) is the resolution of $\mathbf{F}_l$. Because $\mathbf{M}_l \in [0,1]^{B_{\mathrm{sz}} \times 1 \times H_l \times W_l}$, the same spatial coefficient modulates every channel of the corresponding skip feature.

The gated skip and resized decoder tensor are then fused as

$$\mathbf{D}_l = \mathcal{R}_l\Big([\widetilde{\mathbf{D}}_{l+1}, \mathbf{F}_l \odot \mathbf{M}_l]\Big), \qquad \mathbf{D}_5 \equiv \mathbf{F}_5, \tag{32}$$

where $\mathcal{R}_l$ is a convolution, normalization, activation, and residual refinement operator. Equation (32) weights the encoder skip through the gate while passing the deeper decoder stream at full strength.

Applying Eq. (32) from deep to shallow gives

$$\mathbf{D}_4 = \mathcal{G}_4(\mathbf{F}_5, \mathbf{F}_4), \qquad \mathbf{D}_3 = \mathcal{G}_3(\mathbf{D}_4, \mathbf{F}_3), \qquad \mathbf{D}_2 = \mathcal{G}_2(\mathbf{D}_3, \mathbf{F}_2). \tag{33}$$

Equation (33) is the expanded recurrence referenced in the main paper. Each $\mathcal{G}_l$ applies the skip gate and feature fusion in Eqs. (31) and (32).

RCSL injects its original-view structural tensor at the final decoder scale:

$$\mathbf{F}_r = \phi_{\mathrm{reg}}\big([\mathbf{D}_2, \mathbf{Z}^0]\big). \tag{34}$$

Equation (34) reproduces Eq. (3). The decoder feature $\mathbf{D}_2$ and structural feature $\mathbf{Z}^0$ are concatenated at a common spatial resolution. The fusion operator ϕ_{reg} combines their complementary information to form the coarse region representation.

B.4 RCSL RADIANCE-VIEW CONSTRUCTION

RCSL first reverses ImageNet normalization to recover the stored RGB intensity image in $[0, 1]$:

$$I_{b,c,i,j} = \mathrm{clip}_{[0,1]}(d_c x_{b,c,i,j} + m_c). \tag{35}$$

Equation (35) is the inverse of Eq. (28). We apply the radiance perturbations in this intensity space.

Let the per-pixel RGB mean be

$$\mu_c(\mathbf{I})_{b,1,i,j} = \frac{1}{3} \sum_{c=1}^{3} I_{b,c,i,j}. \tag{36}$$

Equation (36) produces the grayscale center used by saturation adjustment. The subscript c on μ_c identifies the channel axis over which the per-pixel mean is computed.

For saturation factor s_{sat}, the adjusted image is

$$\mathbf{I}_{\mathrm{sat}} = \mu_c(\mathbf{I}) + s_{\mathrm{sat}}\big(\mathbf{I} - \mu_c(\mathbf{I})\big). \tag{37}$$

Equation (37) gives the saturation component of Eq. (4). When $s_{\mathrm{sat}} = 1$, the RGB image is unchanged; when $s_{\mathrm{sat}} < 1$, chromatic deviations from the per-pixel mean shrink; when $s_{\mathrm{sat}} > 1$, they expand. The transformation preserves the spatial coordinates.

Exposure and channel response are applied before the gamma transform:

$$\mathbf{I}_e = \mathrm{clip}_{[0,1]}(e\,\mathbf{g} \odot \mathbf{I}_{\mathrm{sat}}), \qquad \mathbf{I}_\gamma = \mathbf{I}_e^\gamma. \tag{38}$$

Equation (38) separates the first part of Eq. (4). The scalar e controls global exposure, the three-vector $\mathbf{g}$ controls channel-dependent response, and the scalar γ models a monotone tone curve.

A low-frequency relay-wall field is generated from a random 4×4 seed:

$$\begin{aligned} \mathbf{U}_{\text{seed}} &\sim \text{Unif}(-1,1)^{B_{\text{sz}}\times 1\times 4\times 4}, \\ \mathbf{W}_{\text{wall}} &= \uparrow_{(H,W)} \mathbf{U}_{\text{seed}}, \\ \mathbf{J} &= \mathbf{I}_{\gamma} \odot (1 + \rho \mathbf{W}_{\text{wall}}) . \end{aligned} \tag{39}$$

Equation (39) completes Eq. (4). Bilinear enlargement makes $\mathbf{W}_{\text{wall}}$ spatially smooth, while the amplitude ρ controls the strength of the simulated relay-wall response. The low-frequency field applies the same spatial modulation to the three RGB channels.

Let $\boldsymbol{\epsilon}$ be elementwise standard Gaussian noise and define the noisy image by

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{1}), \qquad \mathbf{K} = \mathbf{J} + \eta \boldsymbol{\epsilon}. \tag{40}$$

Equation (40) uses η as the sampled noise standard deviation.

The final radiance view mixes the noisy image with its 3×3 average-pooled version and then renormalizes it:

$$\widetilde{\mathbf{x}} = \text{Norm}_{\text{IM}}\Big(\text{clip}_{[0,1]}((1-\beta)\mathbf{K} + \beta \mathcal{A}_3(\mathbf{K}))\Big) . \tag{41}$$

Equation (41) is Eq. (5). The coefficient β continuously interpolates between the noisy image and a mildly blurred image, while Norm_{IM} reapplies the fixed normalization in Eq. (28). Average pooling mixes neighboring intensities on the same spatial coordinate grid.

The independent sampling ranges are

$$\begin{aligned} &e \sim \text{Unif}(0.75, 1.25), \quad g_c \sim \text{Unif}(0.85, 1.15), \quad s_{\text{sat}} \sim \text{Unif}(0.70, 1.30), \\ &\gamma \sim \text{Unif}(0.85, 1.15), \quad \rho \sim \text{Unif}(0.06, 0.18), \quad \eta \sim \text{Unif}(0, 0.015), \\ &\beta \sim \text{Unif}(0, 0.40). \end{aligned} \tag{42}$$

We sample the acquisition factors independently using the ranges in Eq. (42), with a separate set of factors for each image.

For sample b, the eight normalized acquisition factors and their mini-batch matrix are

$$\begin{aligned} \boldsymbol{\zeta}_{\text{pho},b} &= \left[\frac{e_b - 1}{0.25}, \frac{g_{b,1} - 1}{0.15}, \frac{g_{b,2} - 1}{0.15}, \frac{g_{b,3} - 1}{0.15}, \frac{\gamma_b - 1}{0.15}\right]^{\top} \in \mathbb{R}^5, \\ \boldsymbol{\zeta}_{\text{rad},b} &= \left[\frac{\rho_b}{0.18}, \frac{\eta_b}{0.015}, \frac{\beta_b}{0.40}\right]^{\top} \in \mathbb{R}^3, \\ \boldsymbol{\zeta}_{\text{fac},b} &= \left[\boldsymbol{\zeta}_{\text{pho},b}^{\top}, \boldsymbol{\zeta}_{\text{rad},b}^{\top}\right]^{\top} \in \mathbb{R}^8, \\ \mathbf{Q}_{\text{fac}} &= \left[\boldsymbol{\zeta}_{\text{fac},1}^{\top}; \ldots; \boldsymbol{\zeta}_{\text{fac},B_{\text{sz}}}^{\top}\right] \in \mathbb{R}^{B_{\text{sz}}\times 8}. \end{aligned} \tag{43}$$

Each row of $\mathbf{Q}_{\text{fac}}$ contains exposure, three channel gains, gamma, wall-field amplitude, noise level, and blur strength. Saturation additionally varies during view construction; the predictor target comprises the eight factors listed above.

B.5 RCSL STRUCTURAL REPRESENTATION

Both views pass through the same encoder. For each semantic level $l \in \{2, 3, 4, 5\}$, RCSL projects the raw encoder tensor to C_{str} channels:

$$\mathbf{L}_l^v = \pi_l(\mathbf{E}_l^v) \in \mathbb{R}^{B_{\text{sz}}\times C_{\text{str}}\times H_l\times W_l}, \qquad v \in \{0, 1\}. \tag{44}$$

Equation (44) expands the first part of Eq. (6). Each π_l contains a pointwise channel projection followed by residual convolutional refinement. Parameters are shared between views because the same RCSL module processes $\mathbf{x}$ and $\widetilde{\mathbf{x}}$.

The deepest projected map produces a global token:

$$\mathbf{t}^v = h(\mathrm{GAP}(\mathbf{L}_5^v)) \in \mathbb{R}^{B_{\mathrm{sz}} \times C_{\mathrm{str}}}. \tag{45}$$

Equation (45) expands the second part of Eq. (6). The projector h consists of a linear layer, layer normalization, GELU, and a second linear layer; consequently, $\mathbf{t}^v$ summarizes scene-level information while retaining C_{str} learnable channels.

All four local maps are aligned to the resolution of $\mathbf{L}_2^v$ and fused:

$$\mathbf{S}^v = \phi_{\mathrm{str}}\big([\mathbf{L}_2^v, \uparrow_{(H_2,W_2)} \mathbf{L}_3^v, \uparrow_{(H_2,W_2)} \mathbf{L}_4^v, \uparrow_{(H_2,W_2)} \mathbf{L}_5^v]\big)\,. \tag{46}$$

Equation (46) is Eq. (7) with the resize targets shown explicitly. The output $\mathbf{S}^v$ has C_{str} channels and spatial resolution (H_2, W_2). It combines shallow spatial detail with deeper semantic context at co-located wall positions.

The token is transformed into a channel gate and broadcast spatially:

$$\mathbf{m}_{\mathrm{tok}}^v = \sigma(g_{\mathrm{tok}}(\mathbf{t}^v))\,, \qquad \mathbf{T}_{\mathrm{tok}}^v = \mathrm{Broadcast}_{(H_2,W_2)}(\mathbf{t}^v). \tag{47}$$

Equation (47) defines two distinct tensors. The gate $\mathbf{m}_{\mathrm{tok}}^v$ contains channel coefficients in $(0, 1)$, while $\mathbf{T}_{\mathrm{tok}}^v$ copies the global token to every spatial coordinate through parameter-free broadcasting.

The dense structural representation is

$$\mathbf{Z}^v = \mathbf{S}^v \odot \big(0.5 + \mathrm{Broadcast}_{(H_2,W_2)}(\mathbf{m}_{\mathrm{tok}}^v)\big) + 0.1\mathbf{T}_{\mathrm{tok}}^v. \tag{48}$$

Equation (48) expands Eq. (8). The multiplicative coefficient lies in $(0.5, 1.5)$, so the global token can attenuate or amplify a structural channel while retaining a strictly positive multiplicative weight. The additive term preserves a direct scene-level path with fixed scale 0.1.

B.6 RCSL OBJECTIVE AND ITS DERIVATION

B.6.1 GLOBAL INVARIANCE

For each sample b, normalize the two global tokens by

$$\bar{\mathbf{t}}_b^v = \frac{\mathbf{t}_b^v}{\max(\|\mathbf{t}_b^v\|_2, \varepsilon_{\mathrm{g}})}, \qquad \varepsilon_{\mathrm{g}} = 10^{-12}. \tag{49}$$

Equation (49) normalizes each token across channels. The constant ε_{g} prevents division by zero.

The batch- and channel-averaged global consistency loss is

$$\mathcal{L}_{\mathrm{g}} = \frac{1}{B_{\mathrm{sz}} C_{\mathrm{str}}} \sum_{b=1}^{B_{\mathrm{sz}}} \left\| \bar{\mathbf{t}}_b^0 - \bar{\mathbf{t}}_b^1 \right\|_2^2. \tag{50}$$

Equation (50) uses the same reduction as Eq. (9). It aligns scene-level encodings of the same wall observation under two radiance conditions.

For tokens whose norms exceed the numerical stabilizer, the unaveraged squared distance satisfies

$$\left\| \bar{\mathbf{t}}_b^0 - \bar{\mathbf{t}}_b^1 \right\|_2^2 = 2 - 2\left\langle \bar{\mathbf{t}}_b^0, \bar{\mathbf{t}}_b^1 \right\rangle. \tag{51}$$

Equation (51) follows by expanding the square and using unit norms. For these nondegenerate tokens, minimizing $\mathcal{L}_{\mathrm{g}}$ maximizes cosine agreement.

B.6.2 CO-LOCATED LOCAL INVARIANCE

At encoder level l, channel normalization at sample b and wall coordinate (i, j) is

$$\mathrm{norm}_c(\mathbf{L}_l^v)_{b,:,i,j} = \frac{\mathbf{L}^v_{l,b,:,i,j}}{\max(\|\mathbf{L}^v_{l,b,:,i,j}\|_2, \varepsilon_1)}, \qquad \varepsilon_1 = 10^{-12}. \tag{52}$$

Equation (52) normalizes only across channels. Position-preserving view generation pairs the same wall coordinate (i, j) across the two views.

The scalar SmoothL1 penalty used below is

$$\mathrm{Huber}_1(u_{\mathrm{SL1}}) = \begin{cases} \frac{1}{2}u^2_{\mathrm{SL1}}, & |u_{\mathrm{SL1}}| < 1, \\ |u_{\mathrm{SL1}}| - \frac{1}{2}, & |u_{\mathrm{SL1}}| \geq 1. \end{cases} \tag{53}$$

Equation (53) is quadratic for small discrepancies and linear for large discrepancies. This reduces the influence of a small number of strongly changed feature entries while retaining smooth gradients near zero.

Writing $N_l = B_{\mathrm{sz}} C_{\mathrm{str}} H_l W_l$, the exact local loss is

$$\mathcal{L}_{\mathrm{l}} = \frac{1}{4}\sum_{l=2}^{5} \frac{1}{N_l} \sum_{b,c,i,j} \mathrm{Huber}_1\big(\mathrm{norm}_c(\mathbf{L}_l^0)_{b,c,i,j} - \mathrm{norm}_c(\mathbf{L}_l^1)_{b,c,i,j}\big)\,. \tag{54}$$

Equation (54) expands Eq. (10). The factor $1/4$ gives equal importance to the four semantic levels, independent of their spatial resolutions.

B.6.3 DENSE AFFINITY CONSISTENCY

Let $(H_{\mathrm{aff}}, W_{\mathrm{aff}})$ be the adaptive-pooling grid, with each dimension capped at 16, and set $N_{\mathrm{aff}} = H_{\mathrm{aff}} W_{\mathrm{aff}}$. Pooling, flattening, and descriptor normalization give

$$\mathbf{V}^v_{\mathrm{aff},b,n,:} = \frac{\mathrm{vec}_n\big(\mathrm{AdaPool}_{(H_{\mathrm{aff}},W_{\mathrm{aff}})}(\mathbf{Z}_b^v)\big)}{\max\Big(\big\|\mathrm{vec}_n\big(\mathrm{AdaPool}_{(H_{\mathrm{aff}},W_{\mathrm{aff}})}(\mathbf{Z}_b^v)\big)\big\|_2, \varepsilon_{\mathrm{aff}}\Big)}, \quad \varepsilon_{\mathrm{aff}} = 10^{-12}. \tag{55}$$

Equation (55) defines the normalized descriptors used only for affinity consistency. The index $n \in \{1, \ldots, N_{\mathrm{aff}}\}$ enumerates pooled wall positions, and each descriptor has C_{str} channels.

The dense self-affinity matrix is

$$\mathbf{A}^v_{\mathrm{aff},b} = \mathbf{V}^v_{\mathrm{aff},b}(\mathbf{V}^v_{\mathrm{aff},b})^\top, \qquad [\mathbf{A}^v_{\mathrm{aff},b}]_{n_1,n_2} = \big\langle \mathbf{V}^v_{\mathrm{aff},b,n_1,:}, \mathbf{V}^v_{\mathrm{aff},b,n_2,:} \big\rangle\,. \tag{56}$$

For descriptors with norms above the numerical stabilizer, the matrix in Eq. (56) records pairwise cosine similarities among wall-position descriptors.

The affinity objective is

$$\mathcal{L}_{\mathrm{aff}} = \frac{1}{B_{\mathrm{sz}} N^2_{\mathrm{aff}}} \sum_{b=1}^{B_{\mathrm{sz}}} \big\|\mathbf{A}^0_{\mathrm{aff},b} - \mathbf{A}^1_{\mathrm{aff},b}\big\|_F^2\,. \tag{57}$$

Equation (57) uses the same batch-averaged reduction as Eq. (11). It preserves the pairwise geometry among pooled wall positions even when exposure and other acquisition factors alter feature magnitudes.

B.6.4 VARIANCE FLOOR AND COVARIANCE DECORRELATION

For statistical regularization, each view pools $\mathbf{L}_2^v$, $\mathbf{L}_3^v$, and $\mathbf{L}_4^v$, flattens batch and spatial axes, and appends the global tokens as additional rows:

$$\begin{aligned}
\mathbf{X}^v_{\mathrm{stat},l} &= \mathrm{Rows}(\mathrm{AdaPool}_{16}(\mathbf{L}^v_l))\,, && l \in \{2,3,4\},\\
\mathbf{X}^v_{\mathrm{stat}} &= \mathrm{ConcatRows}\big(\mathbf{X}^v_{\mathrm{stat},2}, \mathbf{X}^v_{\mathrm{stat},3}, \mathbf{X}^v_{\mathrm{stat},4}, \mathbf{t}^v\big)\,, \\
\mathbf{X}^v_{\mathrm{stat}} &\in \mathbb{R}^{M_{\mathrm{stat}} \times C_{\mathrm{str}}}.
\end{aligned} \tag{58}$$

Equation (58) pools the three local feature levels and appends the global tokens. $\mathrm{AdaPool}_{16}$ caps both pooled dimensions at 16 and leaves a smaller dimension unchanged. M_{stat} is the resulting number of descriptor rows.

The channel-centered descriptor matrix is

$$\breve{\mathbf{X}}^v_{\mathrm{stat}} = \mathbf{X}^v_{\mathrm{stat}} - \mathbf{1}_{M_{\mathrm{stat}}} \left(\frac{1}{M_{\mathrm{stat}}} \mathbf{1}^\top_{M_{\mathrm{stat}}} \mathbf{X}^v_{\mathrm{stat}} \right). \tag{59}$$

Equation (59) subtracts one empirical mean from every channel. The all-ones vector $\mathbf{1}_{M_{\mathrm{stat}}}$ is used only to express row broadcasting compactly.

For channel c, the stabilized population standard deviation is

$$\widehat{\varpi}^v_c = \sqrt{\frac{1}{M_{\mathrm{stat}}} \sum_{r=1}^{M_{\mathrm{stat}}} (\breve{X}^v_{\mathrm{stat},r,c})^2 + \varepsilon_{\mathrm{var}}}, \qquad \varepsilon_{\mathrm{var}} = 10^{-4}. \tag{60}$$

Equation (60) computes population variance with normalization $1/M_{\mathrm{stat}}$. The stabilizer $\varepsilon_{\mathrm{var}}$ keeps the square-root derivative finite near zero variance.

With standard-deviation floor $\alpha_{\mathrm{std}} = 0.25$, the per-view variance penalty is

$$\mathcal{L}^v_{\mathrm{var}} = \frac{1}{C_{\mathrm{str}}} \sum_{c=1}^{C_{\mathrm{str}}} \max(0, \alpha_{\mathrm{std}} - \widehat{\varpi}^v_c)\,. \tag{61}$$

Equation (61) is zero for a channel whose empirical standard deviation reaches the floor and positive otherwise. Channels below the standard-deviation floor incur a positive penalty, which decreases as their dispersion increases. At exact descriptor collapse, the penalty is positive and its gradient with respect to the descriptors is zero.

The sample covariance matrix is

$$\mathbf{\Sigma}^v_{\mathrm{stat}} = \frac{(\breve{\mathbf{X}}^v_{\mathrm{stat}})^\top \breve{\mathbf{X}}^v_{\mathrm{stat}}}{M_{\mathrm{stat}} - 1} \in \mathbb{R}^{C_{\mathrm{str}} \times C_{\mathrm{str}}}. \tag{62}$$

Equation (62) uses the sample-covariance denominator $M_{\mathrm{stat}} - 1$. Its diagonal entries measure per-channel dispersion, while its off-diagonal entries measure linear redundancy between distinct channels.

The per-view covariance penalty is

$$\mathcal{L}^v_{\mathrm{cov}} = \frac{1}{C_{\mathrm{str}}(C_{\mathrm{str}} - 1)} \sum_{c_1 \neq c_2} \left(\mathbf{\Sigma}^v_{\mathrm{stat},c_1,c_2}\right)^2. \tag{63}$$

Equation (63) averages only squared off-diagonal entries. Equation (61) supplies the complementary constraint on per-channel dispersion.

With both views enabled, the two statistical terms used by RCSL are

$$\mathcal{L}_{\mathrm{var}} = \tfrac{1}{2}\left(\mathcal{L}^0_{\mathrm{var}} + \mathcal{L}^1_{\mathrm{var}}\right), \qquad \mathcal{L}_{\mathrm{cov}} = \tfrac{1}{2}\left(\mathcal{L}^0_{\mathrm{cov}} + \mathcal{L}^1_{\mathrm{cov}}\right). \tag{64}$$

Equation (64) gives equal weight to the original and perturbed descriptor distributions. Each view therefore contributes its own variance and covariance penalties.

B.6.5 ACQUISITION-FACTOR ADVERSARIAL BRANCH

Let $\mathrm{Pred}_{\mathrm{fac},\psi}$ be the two-layer factor predictor and let GRL_υ denote gradient reversal with scale υ. Its prediction and regression loss are

$$\widehat{\mathbf{Q}}_{\mathrm{fac}} = \mathrm{Pred}_{\mathrm{fac},\psi}\big(\mathrm{GRL}_\upsilon(\mathbf{t}^1)\big)\,, \qquad \mathcal{L}_{\mathrm{adv}} = \frac{1}{8B_{\mathrm{sz}}}\left\|\widehat{\mathbf{Q}}_{\mathrm{fac}} - \mathbf{Q}_{\mathrm{fac}}\right\|_F^2. \tag{65}$$

Equation (65) uses only the perturbed-view token because its sampled acquisition factors are known. The predictor learns to estimate the eight factors in Eq. (43), while the structural encoder receives the reversed gradient described next.

Gradient reversal is the identity in the forward pass and changes the derivative sign in the backward pass:

$$\mathrm{GRL}_\upsilon(\mathbf{v}_{\mathrm{rev}}) = \mathbf{v}_{\mathrm{rev}}, \qquad \frac{\partial\,\mathrm{GRL}_\upsilon}{\partial \mathbf{v}_{\mathrm{rev}}} = -\upsilon\mathbf{I}_{\mathrm{id}}, \qquad \upsilon = 1. \tag{66}$$

Equation (66) uses $\mathbf{I}_{\mathrm{id}}$ only for the identity Jacobian. Therefore, the predictor parameters follow the ordinary descent direction, whereas the representation parameters receive its negative.

Writing θ_{RCSL} for the parameters that produce $\mathbf{t}^1$, define the corresponding ordinary, non-reversed regression objective $\mathcal{J}_{\mathrm{fac}}$. The two gradient directions are

$$\begin{aligned} \mathcal{J}_{\mathrm{fac}} &= \frac{1}{8B_{\mathrm{sz}}}\left\|\mathrm{Pred}_{\mathrm{fac},\psi}(\mathbf{t}^1) - \mathbf{Q}_{\mathrm{fac}}\right\|_F^2, \\ \nabla_\psi \mathcal{L}_{\mathrm{adv}} &= \nabla_\psi \mathcal{J}_{\mathrm{fac}}, \\ \nabla_{\theta_{\mathrm{RCSL}}} \mathcal{L}_{\mathrm{adv}} &= -\upsilon \nabla_{\theta_{\mathrm{RCSL}}} \mathcal{J}_{\mathrm{fac}}. \end{aligned} \tag{67}$$

Equation (67) formalizes the adversarial effect. $\mathrm{Pred}_{\mathrm{fac},\psi}$ becomes a stronger acquisition-factor predictor, while the RCSL representation is optimized to make those factors harder to recover. This adversarial objective discourages encoding of the sampled acquisition factors.

B.6.6 COMPLETE RCSL OBJECTIVE

Combining the five mechanisms yields

$$\mathcal{L}_{\mathrm{RCSL}} = \mathcal{L}_{\mathrm{g}} + \mathcal{L}_{\mathrm{l}} + 0.10\,(\mathcal{L}_{\mathrm{var}} + \mathcal{L}_{\mathrm{cov}}) + 0.25\mathcal{L}_{\mathrm{aff}} + 0.10\mathcal{L}_{\mathrm{adv}}. \tag{68}$$

Equation (68) is Eq. (12) with every constituent term defined. Global and local losses align corresponding content, the affinity loss aligns spatial relations, variance and covariance losses maintain informative nonredundant channels, and the adversarial term discourages encoding of known acquisition factors.

B.7 UGBR BOUNDARY AND REGION DERIVATION

B.7.1 COARSE REGION ESTIMATE

UGBR first refines the fused decoder tensor and predicts a coarse full-resolution logit:

$$\mathbf{R} = \phi_R(\mathbf{F}_r), \qquad \mathbf{z}_R = \uparrow_{(H,W)} w_R(\mathbf{R}), \qquad \mathbf{P}_R = \sigma(\mathbf{z}_R). \tag{69}$$

Equation (69) expands Eq. (13). The tensor $\mathbf{R}$ carries region context, $\mathbf{z}_R$ contains unconstrained foreground logits, and $\mathbf{P}_R$ contains the corresponding coarse foreground probabilities.

The boundary stem combines shallow detail, region context, and RCSL structure at the resolution of the shallow detail tensor:

$$\begin{aligned}\mathbf{G}_{\mathrm{det}} &= \pi_d(\mathbf{F}_1),\\ \mathbf{G}_{\mathrm{aux}} &= \pi_{\mathrm{aux}}(\mathbf{Z}^0),\\ \mathbf{B} &= \phi_B\big([\mathbf{G}_{\mathrm{det}}, \uparrow_{(H_1,W_1)} \mathbf{R}, \uparrow_{(H_1,W_1)} \mathbf{G}_{\mathrm{aux}}]\big)\,.\end{aligned} \tag{70}$$

Equation (70) expands the feature construction in Eq. (14). $\mathbf{G}_{\mathrm{det}}$ contributes shallow spatial detail, the resized $\mathbf{R}$ contributes region semantics, and the resized $\mathbf{G}_{\mathrm{aux}}$ contributes radiance-consistent structure.

Three residual branches use dilation factors $\delta_1 = 1$, $\delta_2 = 2$, and $\delta_3 = 3$:

$$\mathbf{H}_B^k = \mathcal{B}_k(\mathbf{B}; \delta_k), \qquad \mathbf{z}_B^k = \uparrow_{(H,W)} w_B^k(\mathbf{H}_B^k), \qquad k \in \{1,2,3\}. \tag{71}$$

Equation (71) defines one feature and one logit map at each boundary scale. The three predictors have separate parameters. The separately parameterized branches learn agreement through shared evidence and joint supervision.

The consensus boundary logit and probability are

$$\overline{\mathbf{z}}_B = \frac{1}{3}\sum_{k=1}^{3} \mathbf{z}_B^k, \qquad \mathbf{P}_B = \sigma(\overline{\mathbf{z}}_B). \tag{72}$$

Equation (72) completes Eq. (14). Averaging is performed in logit space, after which one sigmoid yields the supervised consensus probability.

B.7.2 Boundary disagreement and confidence

For convenience, define the branch probabilities by

$$\mathbf{P}_B^k = \sigma(\mathbf{z}_B^k), \qquad \mathbf{P}_{B,\mathrm{avg}} = \frac{1}{3}\sum_{k=1}^{3} \mathbf{P}_B^k. \tag{73}$$

Equation (73) introduces the probability maps used only to estimate disagreement. $\mathbf{P}_{B,\mathrm{avg}}$ averages the branch probabilities, whereas $\mathbf{P}_B$ applies the sigmoid to the mean logit; these operations generally yield different values.

The population standard deviation across scales is

$$\mathbf{D}_B = \sqrt{\frac{1}{3}\sum_{k=1}^{3} \big(\mathbf{P}_B^k - \mathbf{P}_{B,\mathrm{avg}}\big)^2}. \tag{74}$$

Equation (74) is the first part of Eq. (15). The standard deviation is computed across the three branch probabilities at each pixel.

The detached confidence map is

$$\mathbf{C}_B = \exp(-4\,\mathrm{stop}(\mathbf{D}_B))\,. \tag{75}$$

Equation (75) is the second part of Eq. (15). Agreement gives $\mathbf{D}_B = 0$ and hence $\mathbf{C}_B = 1$, whereas disagreement attenuates the later residual correction.

For three probabilities in $[0,1]$, the disagreement and confidence obey

$$
\begin{aligned}
0 \le D_{B,b,1,i,j} \le \frac{\sqrt{2}}{3}, \\
\exp\left(-\frac{4\sqrt{2}}{3}\right) \le C_{B,b,1,i,j} \le 1, \\
g(d) = \exp(-4d), \qquad g'(d) = -4\exp(-4d) < 0, \\
C_{B,b,1,i,j} = g(\mathrm{stop}(D_{B,b,1,i,j}))\,.
\end{aligned}
\tag{76}
$$

The bounds in Eq. (76) follow from the largest population variance of three values in $[0, 1]$, attained when one value is at one endpoint and the other two are at the opposite endpoint. The decreasing scalar function g converts stronger disagreement into lower confidence. During backpropagation, the stop-gradient operator treats $\mathbf{D}_B$ as constant along this weighting path.

B.7.3 Uncertainty band

The uncertainty of the coarse region posterior is

$$
\mathbf{U}_R = 1 - |2\mathbf{P}_R - 1| = 2\min(\mathbf{P}_R, 1 - \mathbf{P}_R). \tag{77}
$$

Equation (77) expands the first part of Eq. (16). It equals one at probability 0.5 and zero at probabilities zero or one, so $\mathbf{U}_R$ measures uncertainty in coarse foreground membership.

The continuous band gate is

$$
\mathbf{A}_R = \sigma\left(\frac{\mathbf{U}_R - \tau_{\mathrm{band}}}{T_{\mathrm{band}}}\right), \qquad \tau_{\mathrm{band}} = 0.45, \qquad T_{\mathrm{band}} = 0.08. \tag{78}
$$

Equation (78) completes Eq. (16). τ_{band} sets the uncertainty level at which the gate equals 0.5, and T_{band} controls the softness of the transition.

For any positive temperature, the interval on which the band gate exceeds 0.5 is

$$
\begin{aligned}
A_R > \tfrac{1}{2} \quad &\Longleftrightarrow \quad U_R > \tau_{\mathrm{band}} \\
&\Longleftrightarrow \quad \frac{\tau_{\mathrm{band}}}{2} < P_R < 1 - \frac{\tau_{\mathrm{band}}}{2} \\
&\Longleftrightarrow \quad 0.225 < P_R < 0.775.
\end{aligned}
\tag{79}
$$

Equation (79) follows by substituting the second form in Eq. (77). The finite temperature gives a continuous weight both inside and outside this interval; the interval identifies where the uncertainty gate exceeds one half.

B.7.4 Boundary and region relational correction

The exact full-resolution relation input is

$$
\mathbf{H}_{\mathrm{rel}} = \left[\uparrow_{(H,W)} \mathbf{R}, \uparrow_{(H,W)} \mathbf{B}, \uparrow_{(H,W)} \mathbf{G}_{\mathrm{aux}}, \mathbf{P}_R, \mathbf{P}_B, \mathbf{A}_R\right]. \tag{80}
$$

All six relation inputs in Eq. (80) are aligned at the output resolution and concatenated channel-wise. The first three are learned feature tensors, and the last three are one-channel probability or gating maps.

The relation head predicts one additive logit residual:

$$
\Delta\mathbf{z} = w_{\mathrm{rel}}(\mathbf{H}_{\mathrm{rel}}). \tag{81}
$$

The relation stem and head w_{rel} predict a signed logit residual $\Delta\mathbf{z}$. Its sign and magnitude determine the proposed change to the coarse logit.

The joint correction gate and final logit are

$$\mathbf{G}_{\mathrm{corr}} = \mathbf{A}_R \odot \mathbf{C}_B, \qquad \mathbf{z}_{\mathrm{final}} = \mathbf{z}_R + \mathbf{G}_{\mathrm{corr}} \odot \Delta\mathbf{z}. \tag{82}$$

Equation (82) expands Eq. (17). The region gate identifies ambiguous assignments, while boundary confidence controls how strongly the learned residual modifies them. The gates assign continuous positive weights, including in confident regions; the relation head supplies a signed residual of unrestricted magnitude.

B.8 UGBR SUPERVISION AND TOPOLOGY SURROGATE

B.8.1 BINARY REGION OBJECTIVES

For a scalar logit u_{bce} and binary label v_{bce}, define the elementwise loss and its tensor mean by

$$\begin{aligned} \ell_{\mathrm{bce}}(u_{\mathrm{bce}}, v_{\mathrm{bce}}) &= -v_{\mathrm{bce}} \log \sigma(u_{\mathrm{bce}}) - (1 - v_{\mathrm{bce}}) \log(1 - \sigma(u_{\mathrm{bce}})), \\ \ell_{\mathrm{bce}}(u_{\mathrm{bce}}, v_{\mathrm{bce}}) &= \max(u_{\mathrm{bce}}, 0) - u_{\mathrm{bce}} v_{\mathrm{bce}} + \log(1 + \exp(-|u_{\mathrm{bce}}|)), \\ \mathrm{BCE}(\mathbf{z}, \mathbf{y}) &= \frac{1}{B_{\mathrm{sz}} HW} \sum_{b,i,j} \ell_{\mathrm{bce}}(z_{b,1,i,j}, y_{b,1,i,j}). \end{aligned} \tag{83}$$

Equation (83) gives two algebraically equivalent forms of the elementwise loss and then averages it over the tensor. The second expression evaluates the loss directly from logits for numerical stability.

For a probability map $\mathbf{P}$ and target $\mathbf{y}$, the soft Dice loss is

$$\mathrm{Dice}(\mathbf{P}, \mathbf{y}) = \frac{1}{B_{\mathrm{sz}}} \sum_{b=1}^{B_{\mathrm{sz}}} \left[1 - \frac{2 \sum_{i,j} P_{b,1,i,j} y_{b,1,i,j} + 1}{\sum_{i,j} P_{b,1,i,j} + \sum_{i,j} y_{b,1,i,j} + 1} \right]. \tag{84}$$

Equation (84) measures foreground overlap. The additive constant of one stabilizes the overlap ratio for empty foreground sets.

The coarse region objective is

$$\mathcal{L}_R = \mathrm{BCE}(\mathbf{z}_R, \mathbf{y}) + \mathrm{Dice}(\mathbf{P}_R, \mathbf{y}). \tag{85}$$

Equation (85) defines the coarse-region term used in Eq. (18). BCE supplies pixelwise foreground supervision, while Dice supplies region-level overlap supervision.

B.8.2 BOUNDARY TARGET AND CLASS-BALANCED BOUNDARY OBJECTIVE

For the generic single-channel map $\mathbf{O}_{\mathrm{morph}}$, define dilation, erosion, and the morphological edge operator by

$$\begin{aligned} \mathrm{Dil}_3(\mathbf{O}_{\mathrm{morph}}) &= \mathrm{MaxPool}_3(\mathbf{O}_{\mathrm{morph}}), \\ \mathrm{Ero}_3(\mathbf{O}_{\mathrm{morph}}) &= -\mathrm{MaxPool}_3(-\mathbf{O}_{\mathrm{morph}}), \\ \mathrm{Edge}_3(\mathbf{O}_{\mathrm{morph}}) &= \mathrm{clip}_{[0,1]}(\mathrm{Dil}_3(\mathbf{O}_{\mathrm{morph}}) - \mathrm{Ero}_3(\mathbf{O}_{\mathrm{morph}})). \end{aligned} \tag{86}$$

The morphological gradient in Eq. (86) is the difference between the local maximum and minimum over a 3×3 neighborhood. For a binary mask, the result marks pixels whose local neighborhood contains both foreground and background; for a soft probability map, it provides an almost-everywhere differentiable local range.

The supervised boundary target is

$$\mathbf{y}_B = \mathrm{Edge}_3(\mathbf{y}). \tag{87}$$

Equation (87) defines the boundary target used in Eq. (18). It obtains boundary supervision directly from the existing region annotation.

For sample b, let the positive and negative boundary-pixel counts and their clipped ratio be

$$
\begin{aligned}
N_{+,b} &= \sum_{i,j} y_{B,b,1,i,j}, \\
N_{-,b} &= HW - N_{+,b}, \\
\vartheta_{+,b} &= \mathrm{clip}_{[1,20]}\left(\frac{N_{-,b}}{N_{+,b}+\varepsilon_B}\right), \qquad \varepsilon_B = 10^{-6}.
\end{aligned}
\tag{88}
$$

Equation (88) compensates for sparse positive contours. The lower bound avoids down-weighting positives, and the upper bound prevents a very small contour from producing an unstable loss multiplier.

The pixel weight and weighted boundary BCE are

$$
\begin{aligned}
\vartheta^B_{b,i,j} &= 1 + y_{B,b,1,i,j}(\vartheta_{+,b} - 1), \\
\mathrm{wBCE}(\overline{\mathbf{z}}_B, \mathbf{y}_B) &= \frac{1}{B_{\mathrm{sz}} HW} \sum_{b,i,j} \vartheta^B_{b,i,j}\, \ell_{\mathrm{bce}}(\overline{z}_{B,b,1,i,j}, y_{B,b,1,i,j}).
\end{aligned}
\tag{89}
$$

Equation (89) assigns unit weight to background pixels and $\vartheta_{+,b}$ to boundary pixels. The scalar function ℓ_{bce} is the stable elementwise logit loss defined in Eq. (83).

The full boundary objective is

$$
\mathcal{L}_B = \mathrm{wBCE}(\overline{\mathbf{z}}_B, \mathbf{y}_B) + \mathrm{Dice}(\mathbf{P}_B, \mathbf{y}_B). \tag{90}
$$

Equation (90) defines the boundary term used in Eq. (18). Weighted BCE handles contour sparsity, while Dice directly rewards overlap between the consensus boundary probability and the derived target boundary.

B.8.3 Boundary and region relation objective

Define the mean absolute difference between equal-sized maps by

$$
\mathrm{MAE}(\mathbf{U}_{\mathrm{mae}}, \mathbf{V}_{\mathrm{mae}}) = \frac{1}{B_{\mathrm{sz}} HW} \|\mathbf{U}_{\mathrm{mae}} - \mathbf{V}_{\mathrm{mae}}\|_1. \tag{91}
$$

The MAE operator in Eq. (91) averages absolute differences over the batch and spatial dimensions.

The boundary extracted from the final foreground posterior is

$$
\mathbf{P}_{\mathrm{edge}} = \mathrm{Edge}_3(\sigma(\mathbf{z}_{\mathrm{final}}))\,. \tag{92}
$$

Equation (92) converts the final region probability into the soft contour $\mathbf{P}_{\mathrm{edge}}$ using Eq. (86). It creates the quantity through which region and boundary predictions are related.

The relation objective is

$$
\mathcal{L}_{\mathrm{rel}} = \mathrm{MAE}(\mathbf{P}_B, \mathbf{P}_{\mathrm{edge}}) + 0.5\,\mathrm{MAE}(\mathbf{P}_{\mathrm{edge}}, \mathbf{y}_B). \tag{93}
$$

Equation (93) expands Eq. (19). The first term makes the consensus boundary agree with the boundary implied by the final region; the second anchors that implied boundary to ground truth with half weight.

B.8.4 Scale-normalized soft Euler surrogate

For any probability map $\mathbf{P}$, resize it bilinearly to

$$
H_{\mathfrak{E}} = \min(64, H), \qquad W_{\mathfrak{E}} = \min(64, W), \qquad \mathbf{P}^{\downarrow} = \uparrow_{(H_{\mathfrak{E}}, W_{\mathfrak{E}})} \mathbf{P}, \tag{94}
$$

Equation (94) applies bilinear downsampling when either input dimension exceeds 64 and otherwise keeps that dimension unchanged. The resulting grid has at most 64×64 locations.

Pad $\mathbf{P}^{\downarrow}$ by one zero on every side and denote the four probabilities of padded 2×2 cell (i, j) by

$$p_{i,j}^{00} = P_{i,j}^{\mathrm{pad}}, \qquad p_{i,j}^{01} = P_{i,j+1}^{\mathrm{pad}}, \qquad p_{i,j}^{10} = P_{i+1,j}^{\mathrm{pad}}, \qquad p_{i,j}^{11} = P_{i+1,j+1}^{\mathrm{pad}}. \tag{95}$$

Equation (95) fixes a consistent orientation for every local cell. The cell indices range over $i \in \{1, \ldots, H_{\mathfrak{E}} + 1\}$ and $j \in \{1, \ldots, W_{\mathfrak{E}} + 1\}$. Zero padding declares the exterior to be background, so objects touching an image edge still create boundary-transition patterns in the local configuration counts.

Under independent Bernoulli occupancy with probabilities $p_{i,j}^{ab}$, the probability that exactly one corner of cell (i, j) is foreground is

$$\begin{aligned}\omega_{1,i,j} = {} & p_{i,j}^{00}(1 - p_{i,j}^{01})(1 - p_{i,j}^{10})(1 - p_{i,j}^{11}) + p_{i,j}^{01}(1 - p_{i,j}^{00})(1 - p_{i,j}^{10})(1 - p_{i,j}^{11}) \\ & + p_{i,j}^{10}(1 - p_{i,j}^{00})(1 - p_{i,j}^{01})(1 - p_{i,j}^{11}) + p_{i,j}^{11}(1 - p_{i,j}^{00})(1 - p_{i,j}^{01})(1 - p_{i,j}^{10}),\end{aligned} \tag{96}$$

where all four factors on the right-hand side belong to the same cell (i, j). Equation (96) sums the four mutually exclusive configurations with one occupied corner.

The probability that exactly three corners are foreground is

$$\begin{aligned}\omega_{3,i,j} = {} & (1 - p_{i,j}^{00})p_{i,j}^{01}p_{i,j}^{10}p_{i,j}^{11} + (1 - p_{i,j}^{01})p_{i,j}^{00}p_{i,j}^{10}p_{i,j}^{11} \\ & + (1 - p_{i,j}^{10})p_{i,j}^{00}p_{i,j}^{01}p_{i,j}^{11} + (1 - p_{i,j}^{11})p_{i,j}^{00}p_{i,j}^{01}p_{i,j}^{10}.\end{aligned} \tag{97}$$

Equation (97) sums the four mutually exclusive configurations with one empty corner. These configurations contribute with the opposite sign to the local Euler estimate.

The probability of a two-corner diagonal configuration is

$$\omega_{\times,i,j} = p_{i,j}^{00}p_{i,j}^{11}(1 - p_{i,j}^{01})(1 - p_{i,j}^{10}) + p_{i,j}^{01}p_{i,j}^{10}(1 - p_{i,j}^{00})(1 - p_{i,j}^{11}). \tag{98}$$

Equation (98) contains the two possible diagonals. The two diagonal configurations enter the soft Euler count with coefficient -2.

Summing local configuration probabilities gives the three soft counts

$$N_1^{\mathrm{soft}} = \sum_{i,j} \omega_{1,i,j}, \qquad N_3^{\mathrm{soft}} = \sum_{i,j} \omega_{3,i,j}, \qquad N_{\times}^{\mathrm{soft}} = \sum_{i,j} \omega_{\times,i,j}. \tag{99}$$

Equation (99) replaces hard counts of binary 2×2 patterns by their differentiable expected counts. If every probability is zero or one, these expressions reduce exactly to ordinary pattern counts.

The spatial normalizer and the soft Euler surrogate are

$$\varrho_{\mathfrak{E}} = \sqrt{\max((H_{\mathfrak{E}} - 1)(W_{\mathfrak{E}} - 1), 1)}, \qquad \mathfrak{E}(\mathbf{P}) = \frac{N_1^{\mathrm{soft}} - N_3^{\mathrm{soft}} - 2N_{\times}^{\mathrm{soft}}}{4\varrho_{\mathfrak{E}}}. \tag{100}$$

In Eq. (100), the factor $1/4$ converts local configuration counts into the soft Euler statistic, and $\varrho_{\mathfrak{E}}$ scales it for the auxiliary structural objective. Matching this scalar encourages agreement in the component-and-hole statistic, whose value can be shared by distinct shapes and topologies.

Let $\mathbf{P}_F = \sigma(\mathbf{z}_{\mathrm{final}})$. The topology objective is

$$\mathcal{L}_{\mathrm{top}} = \frac{1}{B_{\mathrm{sz}}} \sum_{b=1}^{B_{\mathrm{sz}}} \mathrm{Huber}_1(\mathfrak{E}(\mathbf{P}_{F,b}) - \mathfrak{E}(\mathbf{y}_b)). \tag{101}$$

Equation (101) expands Eq. (20). It penalizes disagreement in the scale-normalized expected component-and-hole statistic while remaining differentiable with respect to the final probability map.

B.8.5 COMPLETE UGBR OBJECTIVE

The auxiliary objective for UGBR is

$$\mathcal{L}_{\mathrm{UGBR}} = 0.50\mathcal{L}_R + 0.75\mathcal{L}_B + 0.30\mathcal{L}_{\mathrm{rel}} + 0.02\mathcal{L}_{\mathrm{top}}. \tag{102}$$

Equation (102) is Eq. (21). The weights determine the balance within the UGBR auxiliary branch before its outer multiplier is applied.

B.9 DEEP SUPERVISION, COMPLETE TRAINING OBJECTIVE, AND INFERENCE

The three decoder side logits are

$$\begin{aligned} \mathbf{z}^4_{\mathrm{side}} &= \uparrow_{(H,W)} w^4_{\mathrm{side}}(\mathbf{D}_4), \\ \mathbf{z}^3_{\mathrm{side}} &= \uparrow_{(H,W)} w^3_{\mathrm{side}}(\mathbf{D}_3), \\ \mathbf{z}^2_{\mathrm{side}} &= \uparrow_{(H,W)} w^2_{\mathrm{side}}(\mathbf{D}_2). \end{aligned} \tag{103}$$

Equation (103) makes the three side outputs referenced in the main paper explicit. Each side head is a learned 1×1 convolution followed by bilinear resizing to the target resolution.

The indexed collection of supervised logits is

$$\mathcal{S} = \left(\mathbf{z}_{\mathrm{final}}, \mathbf{z}_R, \mathbf{z}^4_{\mathrm{side}}, \mathbf{z}^3_{\mathrm{side}}, \mathbf{z}^2_{\mathrm{side}}\right). \tag{104}$$

Equation (104) contains five supervised entries: the final prediction, the coarse region prediction, and the three decoder predictions. With UGBR enabled but relation correction disabled, the final and coarse predictions coincide and occupy separate entries. With UGBR disabled, the coarse entry is omitted.

The soft intersection-over-union term is

$$\begin{aligned} \iota_{\cap,b}(\mathbf{P},\mathbf{y}) &= \sum_{i,j} P_{b,1,i,j} y_{b,1,i,j}, \\ \iota_{\cup,b}(\mathbf{P},\mathbf{y}) &= \sum_{i,j} \left(P_{b,1,i,j} + y_{b,1,i,j} - P_{b,1,i,j} y_{b,1,i,j}\right), \\ \mathcal{L}_{\mathrm{IoU}}(\mathbf{P},\mathbf{y}) &= \frac{1}{B_{\mathrm{sz}}} \sum_{b=1}^{B_{\mathrm{sz}}} \left[1 - \frac{\iota_{\cap,b}(\mathbf{P},\mathbf{y}) + \varepsilon_{\mathrm{IoU}}}{\iota_{\cup,b}(\mathbf{P},\mathbf{y}) + \varepsilon_{\mathrm{IoU}}}\right], \qquad \varepsilon_{\mathrm{IoU}} > 0. \end{aligned} \tag{105}$$

Equation (105) first defines the soft intersection $\iota_{\cap,b}$ and union $\iota_{\cup,b}$, then forms the differentiable IoU complement used in Eq. (22). The positive numerical stabilizer keeps the ratio defined for an empty union.

Deep segmentation supervision is

$$\mathcal{L}_{\mathrm{seg}} = \sum_{k=1}^{|\mathcal{S}|} \left[\mathrm{BCE}(\mathcal{S}_k, \mathbf{y}) + \mathcal{L}_{\mathrm{IoU}}(\sigma(\mathcal{S}_k), \mathbf{y})\right]. \tag{106}$$

Equation (106) is Eq. (22). For the full model, applying the same region objective to all five entries supplies gradients to intermediate decoder stages and to the UGBR coarse region head.

The total training loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{seg}} + 0.05\mathcal{L}_{\mathrm{RCSL}} + 0.25\mathcal{L}_{\mathrm{UGBR}}. \tag{107}$$

Equation (107) is Eq. (23). The internal RCSL and UGBR weights are applied first, and the outer coefficients 0.05 and 0.25 then control how strongly the two components regularize the deeply supervised segmentation objective.

At inference, the original-view features and the UGBR correction produce the foreground probability map:

$$\widehat{\mathbf{y}}_{\text{test}} = \sigma(\mathbf{z}_R(\mathbf{x}) + \mathbf{A}_R(\mathbf{x}) \odot \mathbf{C}_B(\mathbf{x}) \odot \Delta\mathbf{z}(\mathbf{x})) \,. \tag{108}$$

At inference, RUBIN uses the original-view structural feature $\mathbf{Z}^0$ and the UGBR correction to produce the foreground probability map in Eq. (108). Radiance-view generation and the auxiliary learning objectives are used during training.

B.10 How the two components cooperate

Radiance-consistent structural learning. RCSL changes acquisition-related intensities while keeping the wall-coordinate grid fixed. It aligns scene tokens, co-located local descriptors, and dense self-affinities. Variance and covariance regularization encourage descriptor diversity and reduce channel redundancy, while gradient reversal discourages encoding of sampled acquisition factors.

Uncertainty-gated refinement. The band gate $\mathbf{A}_R$ weights ambiguous coarse-region assignments, and $\mathbf{C}_B$ summarizes agreement across the boundary branches. Their product weights the relational residual in Eq. (82). This couples the location of refinement to region uncertainty and its strength to multiscale boundary agreement. Region, boundary, relation, and soft Euler objectives provide complementary supervision.

Coupling of the components. The two components meet at two points in the recovery process. First, $\mathbf{Z}^0$ is fused with $\mathbf{D}_2$ to build the coarse region representation. Second, its projection $\mathbf{G}_{\text{aux}}$ supplies structural evidence to both the boundary stem and the relation head. UGBR therefore refines the foreground using evidence that is also available to the coarse estimate. RCSL encourages stable representations of wall light, while UGBR determines where and how boundary evidence reshapes the silhouette.

C Mechanistic Evidence for the Two Components

We examine how RCSL and UGBR support silhouette recovery on the 200-image NLOS-ECSSD test split, with model parameters and diagnostic evaluation settings fixed. The diagnostics follow the recovery process: cross-view similarity and relation drift characterize representation stability, acquisition-factor probes examine radiance encoding, and boundary measurements characterize the evidence available within uncertain regions.

Table 3: Selected mechanism diagnostics on the 200-image NLOS-ECSSD test split. Image-level quantities are averaged over test images; the held-out factor R^2 is averaged over the five modeled photometric factors. Lower affinity error indicates more stable relations, and lower held-out factor R^2 indicates reduced linear predictability.

Component	Diagnostic	Full	Ablation / region	Observed change
RCSL	Global token cosine similarity ↑	0.99998	0.99881 (without global consistency)	Lower mean cosine similarity without global consistency.
RCSL	Dense affinity error ↓	0.0041	0.0435 (without affinity consistency)	90.6% lower relation drift with affinity consistency.
RCSL	Held-out factor linear R^2 ↓	0.414	0.437 (without factor suppression)	Lower predictability of the modeled acquisition factors.
UGBR	Boundary F1 inside the uncertainty band ↑	0.0572	0.0545 (outside band)	Boundary evidence is slightly stronger inside the selected band.

C.1 RCSL: structural stability and radiance decoupling

We examine whether RCSL retains mask-relevant structure while reducing relation drift under a generated radiance view (Figure 7(a,b)).

VICRegL selects local feature pairs using spatial and feature-space proximity (Bardes et al., 2022b), while PooDLe establishes dense correspondence through optical-flow warping (Wang et al., 2025).

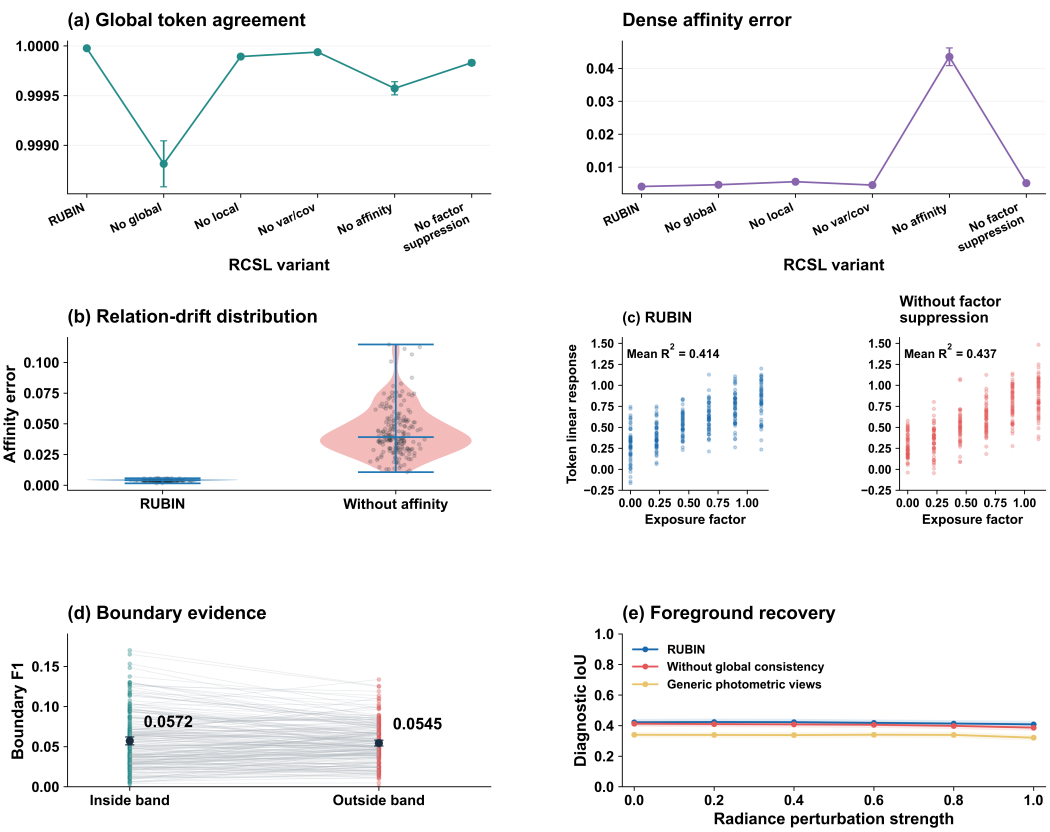

Figure 7: Mechanism diagnostics on the fixed 200-image NLOS-ECSSD test split. (a) RCSL component sensitivity: global token cosine similarity and dense affinity error. (b) Affinity-error distributions with and without affinity consistency. (c) Acquisition-factor leakage illustrated by exposure probes, with mean held-out linear R^2 over the five modeled factors. (d) Boundary F1 inside and outside the UGBR uncertainty band. (e) Diagnostic IoU under six position-preserving radiance perturbation strengths for the full model, the no-global variant, and the photometric-view replacement. ADRSR denotes RCSL in the plot's implementation labels.

RCSL obtains matched wall coordinates directly from position-preserving radiance views. Adding its affinity objective lowers relation drift from 0.0435 to 0.0041 in our ablation study, a reduction of 90.6% (Table 3). The affinity-error distributions in Figure 7(b) are more concentrated near zero for the full model, consistent with more stable spatial relations across radiance views.

Global-token consistency complements these spatial relations, with mean cosine similarity of 0.99998 for RUBIN and 0.99881 for the variant without global consistency. The affinity diagnostic complements these near-unit token similarities by separating the corresponding variants more clearly.

Acquisition-factor suppression lowers mean held-out linear R^2 from 0.437 to 0.414 (Figure 7(c)). The five evaluated photometric factors become less linearly predictable from the representation, consistent with features that emphasize structure shared across radiance views. These five probe targets are a subset of the eight regression targets defined in Eq. (43).

Figure 7(e) connects representation stability to mask recovery: the full model remains above both the no-global variant and the photometric-view replacement across the six evaluated radiance strengths. The recovery advantage persists across the position-preserving radiance transformations used in this experiment.

C.2 UGBR: SELECTIVE BOUNDARY CORRECTION

Figure 7(d) evaluates boundary evidence separately inside and outside the uncertainty band.

The uncertainty band retains boundary evidence for correction, with mean boundary F1 of 0.0572 inside the band and 0.0545 outside it (Table 3). UGBR combines this evidence with the coarse region estimate, while the component ablation in Table 2 measures the resulting improvement in foreground recovery. The small band difference is descriptive evidence and does not establish boundary accuracy on its own.

The scale-normalized soft Euler term serves as an auxiliary structural regularizer, complementing the region, boundary, and relation objectives. The topology diagnostics vary across measures, with some component and Euler errors lower when the auxiliary term is removed. It therefore does not establish a separate topology-preservation advantage.

C.3 Connection to Silhouette Recovery

The diagnostics follow the same recovery chain as RUBIN. Cross-view measurements characterize the stability of spatial relations, boundary measurements describe the evidence available for refinement, and component ablations connect the designs to the final mask. The five-dataset comparisons then evaluate their combined contribution to recovering basic hidden-target shape from wall light. These mechanism measurements describe the evaluated model at one fixed seed on one test split.

D Extended Quantitative and Qualitative Comparisons

D.1 Extended Quantitative Comparison

The main paper reports Max-F, MAE, and Dice as the primary metrics. This appendix adds Mean-F, the weighted F-measure F_β^w, the structure measure S_m, and the enhanced alignment measure E_m. Max-F is the maximum of the F-measure curve formed from test-set average precision and recall over the evaluated thresholds. Mean-F averages per-image F-measures obtained with adaptive thresholds equal to twice the prediction mean, clipped to one. It is therefore an adaptive-threshold score, not a mean over the threshold curve. The weighted F-measure evaluates errors with spatial weighting (Margolin et al., 2014); S_m evaluates object- and region-aware structural similarity (Fan et al., 2017); and E_m combines local agreement with image-level statistics (Fan et al., 2018). Together with foreground overlap and pixelwise MAE, these measures describe complementary properties of the recovered silhouette.

All methods follow the comparison protocol described in Section 4.1. Each model receives the RGB relay-wall observation and predicts a mask in target-image coordinates; the source image provides a visual reference. Tables 4, 5, and 6 report the complete results. Best and second-best values are highlighted separately, with ties assigned the same rank.

Overall performance across the extended metrics. Across all five datasets, RUBIN combines the highest Max-F, Mean-F, weighted F-measure, and S_m with the lowest MAE (Tables 4–6). It also leads Dice on all five datasets. Of the 35 dataset and metric combinations, 33 are uniquely best and one is tied best at the reported precision. The tied value is $E_m = 0.814$ on NLOS-DUTS; on NLOS-DUT-OMRON, RUBIN has $E_m = 0.683$, below LDF at 0.687 and ZoomNeXt at 0.685. The overall improvement therefore spans foreground decisions, spatially weighted accuracy, overlap, and regional structure.

NLOS-HKU. On NLOS-HKU, RUBIN improves foreground recovery under both the best-threshold and image-adaptive rules. The gains extend to spatially weighted accuracy and foreground overlap, with $F_\beta^w = 0.452$ and Dice of 0.497. Together, these results indicate that the recovered region better captures the target's occupied area across complementary criteria.

NLOS-ECSSD. SegFormer combines a hierarchical Transformer with an MLP decoder to integrate multiscale information (Xie et al., 2021). On NLOS-ECSSD, RUBIN achieves $S_m = 0.727$ and $F_\beta^w = 0.663$, compared with 0.551 and 0.405 for SegFormer and the strongest competing values of 0.624 and 0.496, respectively (Table 4). Together with the overlap and pixel-error gains in the main comparison, these results connect more complete foreground coverage with better regional structure.

Table 4: Quantitative comparison on NLOS-HKU and NLOS-ECSSD benchmarks across seven metrics. The best results are highlighted in **red**, and the second-best results in **blue**. Tied values share the same rank.

Method	Venue	NLOS-HKU							NLOS-ECSSD						
		Max-F↑	Mean-F↑	F_β^w ↑	MAE↓	S_m ↑	E_m ↑	Dice↑	Max-F↑	Mean-F↑	F_β^w ↑	MAE↓	S_m ↑	E_m ↑	Dice↑
Direct-view segmentation															
LDF	CVPR'20	**0.527**	**0.504**	**0.409**	0.178	**0.572**	**0.763**	**0.463**	0.604	0.563	0.459	0.194	0.599	0.724	0.517
ZoomNeXt	TPAMI'24	0.520	0.493	0.387	0.178	0.558	0.755	0.434	0.614	0.568	0.485	0.199	0.612	0.720	0.557
GateNet	ECCV'20	0.519	0.496	0.394	0.176	0.562	0.748	0.442	0.621	0.580	**0.496**	0.184	**0.624**	0.738	**0.565**
FAP-Net	TIP'22	0.512	0.486	0.367	0.182	0.552	0.756	0.413	0.605	0.562	0.463	0.191	0.606	0.726	0.525
PGNet	CVPR'22	0.499	0.464	0.366	**0.172**	0.531	0.688	0.391	**0.625**	**0.584**	0.487	**0.175**	0.606	0.737	0.531
BiRefNet	CAAI AIR'24	0.508	0.477	0.378	0.173	0.543	0.719	0.411	0.586	0.538	0.444	0.199	0.573	0.709	0.497
SegFormer	NeurIPS'21	0.466	0.427	0.317	0.192	0.512	0.694	0.350	0.561	0.508	0.405	0.209	0.551	0.674	0.452
CamoFormer	TPAMI'24	0.489	0.449	0.310	0.193	0.518	0.731	0.346	0.581	0.540	0.428	0.200	0.580	0.714	0.476
VSCode	CVPR'24	0.473	0.461	0.365	0.198	0.553	0.711	0.423	0.571	0.531	0.447	0.205	0.589	0.700	0.514
Indirect-imaging methods															
NLOS-OT	TIP'22	0.509	0.479	0.392	0.175	0.547	0.735	0.432	0.601	0.567	0.482	0.184	0.597	0.747	0.541
Learnable Inv. Kernel	ICCV'23	0.505	0.476	0.388	0.176	0.544	0.739	0.428	0.602	0.566	0.477	0.178	0.597	0.744	0.524
NeTF	TPAMI'21	0.511	0.483	0.392	**0.172**	0.546	0.734	0.425	0.600	0.563	0.471	0.182	0.589	0.738	0.515
NLOS-LTM	TIP'25	0.490	0.452	0.343	0.175	0.527	0.674	0.369	0.607	0.559	0.459	0.182	0.589	0.717	0.502
PAC-Net	CVPR'23	0.469	0.435	0.328	0.186	0.516	0.692	0.357	0.584	0.542	0.442	0.188	0.574	0.702	0.483
NLOST	CVPR'23	0.522	0.494	0.406	**0.172**	0.555	0.744	0.447	0.590	0.558	0.476	0.185	0.594	**0.748**	0.533
LEAP	ECCV'24	0.496	0.469	0.363	0.191	0.553	0.736	0.417	0.602	0.562	0.469	0.191	0.608	0.727	0.534
NLOS-LPP	ICCV'25	0.498	0.471	0.363	0.192	0.551	0.741	0.417	0.590	0.553	0.453	0.191	0.600	0.711	0.514
FMoE-NLOS	PR'26	0.492	0.471	0.366	0.191	0.554	0.737	0.419	0.600	0.558	0.460	0.191	0.603	0.722	0.522
ST-Mamba	NeurIPS'24	0.500	0.475	0.367	0.192	0.554	0.748	0.420	0.601	0.559	0.461	0.190	0.603	0.725	0.523
TransiT	ICCV'25	0.489	0.468	0.361	0.191	0.552	0.732	0.414	0.593	0.553	0.457	0.189	0.602	0.714	0.518
Ours															
RUBIN	**Ours**	**0.565**	**0.537**	**0.452**	**0.161**	**0.586**	**0.765**	**0.497**	**0.751**	**0.732**	**0.663**	**0.109**	**0.727**	**0.838**	**0.697**

NLOS-DUTS. RUBIN achieves the best or tied-best value for all seven metrics on NLOS-DUTS. Foreground detection, overlap, and structural agreement improve together, with an E_m of 0.814 shared with ZoomNeXt. This metric profile extends the evidence for coherent foreground recovery to another target collection.

NLOS-DUT-OMRON. On NLOS-DUT-OMRON, RUBIN improves the strongest competing Dice score from 0.340 to 0.361 and weighted F-measure from 0.307 to 0.332. The gain concerns both foreground overlap and the spatial distribution of errors. RUBIN has a lower E_m than LDF and ZoomNeXt, illustrating that enhanced alignment and foreground overlap capture different aspects of the prediction. The complete metric profile is given in Table 5.

NLOS-PASCAL. On NLOS-PASCAL, RUBIN leads all seven reported metrics, including Max-F of 0.488 and weighted F-measure of 0.349. Its Dice score is 0.410, followed by PAC-Net at 0.409 and NLOS-LPP and FMoE-NLOS at 0.408. These results support improved recovery of the target's occupied region and spatial arrangement, with several margins remaining small at the reported precision.

Comparison across method families. The strongest competing method changes across datasets and metrics. LDF is strong on several NLOS-HKU columns; PGNet and GateNet are competitive on NLOS-ECSSD; NLOST, ZoomNeXt, and GateNet are strong on NLOS-DUTS; and other methods provide the closest alternatives on NLOS-DUT-OMRON and NLOS-PASCAL. RUBIN maintains its overall advantage across comparison architectures whose strengths vary with the dataset.

NLOS-OT formulates passive image recovery through manifold mapping and optimal transport (Geng et al., 2022), while NLOS-LTM conditions reconstruction on learned light transport (Zhang et al., 2025). For the RGB-to-mask adaptations in this comparison, RUBIN raises NLOS-ECSSD Dice to 0.697, compared with 0.541 and 0.502 for these two architectures, respectively (Table 4). Together with the component ablations, this comparison supports learning stable structural evidence and using it to form coherent target regions.

D.2 Extended Qualitative Comparison

Figures 8, 9, 10, 11, and 12 present five parts of the extended qualitative comparison alongside the corresponding observations. In every figure, column (1) is the target source image for visual reference, column (2) is the relay wall observation supplied to the models, columns (3)–(22) are

Table 5: Quantitative comparison on NLOS-DUTS and NLOS-DUT-OMRON benchmarks across seven metrics. The best results are highlighted in **red**, and the second-best results in **blue**. Tied values share the same rank.

Method	Venue	NLOS-DUTS							NLOS-DUT-OMRON						
		Max-F↑	Mean-F↑	F_β^w ↑	MAE↓	S_m ↑	E_m ↑	Dice↑	Max-F↑	Mean-F↑	F_β^w ↑	MAE↓	S_m ↑	E_m ↑	Dice↑
Direct-view segmentation															
LDF	CVPR'20	0.651	0.620	0.545	0.151	0.662	0.795	0.612	0.420	0.378	0.280	0.147	0.535	**0.687**	0.315
ZoomNeXt	TPAMI'24	0.674	0.647	0.569	0.143	**0.679**	**0.814**	0.630	0.422	0.384	0.294	0.148	0.544	**0.685**	0.333
GateNet	ECCV'20	0.673	0.639	0.569	0.145	**0.679**	**0.811**	**0.639**	**0.434**	**0.394**	0.303	0.145	**0.546**	0.669	**0.340**
FAP-Net	TIP'22	0.671	0.637	0.560	0.146	0.675	0.807	0.627	0.393	0.348	0.235	0.146	0.516	0.660	0.261
PGNet	CVPR'22	0.619	0.592	0.516	0.160	0.629	0.785	0.573	0.188	0.184	0.001	0.151	0.426	0.250	0.000
BiRefNet	CAAI AIR'24	0.632	0.604	0.531	0.150	0.644	0.773	0.586	0.188	0.184	0.001	0.151	0.426	0.250	0.000
SegFormer	NeurIPS'21	0.571	0.574	0.476	0.154	0.620	0.724	0.523	0.188	0.243	0.107	0.144	0.466	0.397	0.115
CamoFormer	TPAMI'24	0.614	0.577	0.504	0.178	0.642	0.765	0.591	0.378	0.318	0.202	0.160	0.488	0.659	0.209
VSCode	CVPR'24	0.614	0.583	0.502	0.162	0.632	0.743	0.559	0.399	0.363	0.266	0.154	0.526	0.672	0.301
Indirect-imaging methods															
NLOS-OT	TIP'22	0.670	0.645	0.568	**0.137**	0.667	0.807	0.618	0.423	0.375	0.301	**0.137**	0.536	0.665	0.328
Learnable Inv. Kernel	ICCV'23	0.660	0.635	0.558	0.142	0.660	0.803	0.610	0.428	0.380	**0.307**	0.139	0.536	0.671	0.335
NeTF	TPAMI'21	0.673	0.648	0.569	**0.137**	0.666	0.805	0.615	0.415	0.368	0.295	0.140	0.531	0.672	0.322
NLOS-LTM	TIP'25	0.635	0.611	0.535	0.156	0.645	0.800	0.598	0.188	0.184	0.001	0.151	0.426	0.250	0.000
PAC-Net	CVPR'23	0.614	0.589	0.519	0.164	0.635	0.780	0.588	0.277	0.220	0.135	0.148	0.459	0.497	0.138
NLOST	CVPR'23	**0.676**	**0.650**	**0.572**	**0.137**	0.669	0.806	0.621	0.387	0.335	0.263	0.139	0.517	0.639	0.284
LEAP	ECCV'24	0.621	0.583	0.505	0.173	0.644	0.768	0.589	0.403	0.352	0.230	0.160	0.509	0.657	0.260
NLOS-LPP	ICCV'25	0.626	0.588	0.510	0.170	0.646	0.769	0.591	0.402	0.357	0.234	0.161	0.511	0.665	0.266
FMoE-NLOS	PR'26	0.620	0.582	0.502	0.171	0.643	0.761	0.583	0.405	0.358	0.236	0.160	0.513	0.664	0.269
ST-Mamba	NeurIPS'24	0.625	0.588	0.515	0.170	0.649	0.777	0.597	0.410	0.361	0.237	0.161	0.513	0.673	0.270
TransiT	ICCV'25	0.624	0.584	0.508	0.174	0.646	0.770	0.594	0.388	0.348	0.226	0.161	0.508	0.647	0.258
Ours															
RUBIN	**Ours**	**0.692**	**0.662**	**0.584**	**0.136**	**0.683**	**0.814**	**0.640**	**0.447**	**0.403**	**0.332**	**0.134**	**0.552**	0.683	**0.361**

the twenty baselines in the order of Figure 3, column (23) is RUBIN, and column (24) is the target ground truth. Only column (2) is used as input to the evaluated models.

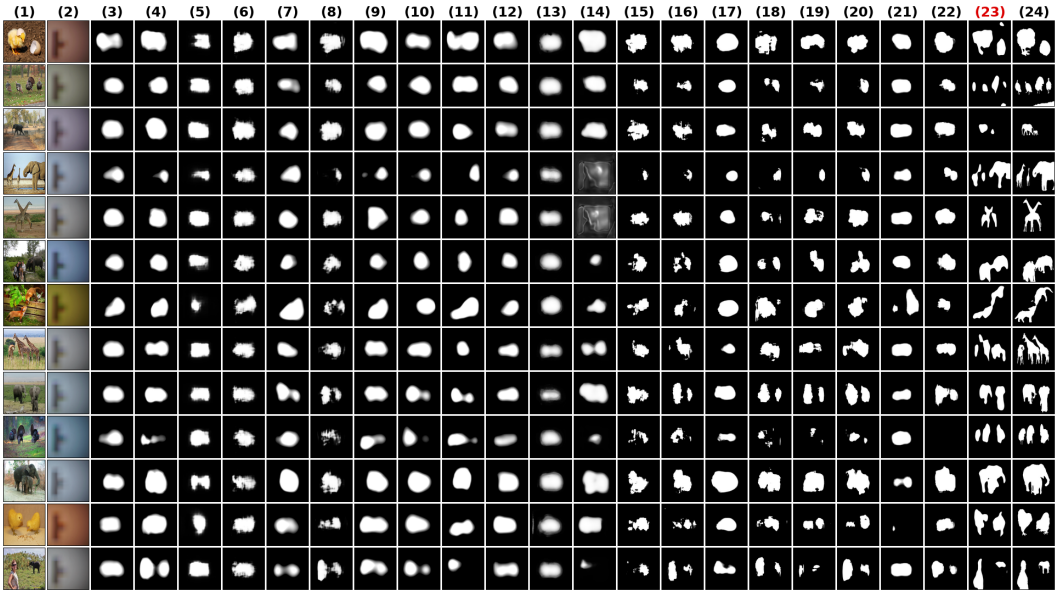

Figure 8: Extended qualitative comparison on measured NLOS observations, Part 1 of 5. The 24-column order follows Figure 3; RUBIN and ground truth occupy the last two columns. All evaluated models receive only the relay wall observation in column (2).

Recovery of the dominant foreground region. The relay wall observations in column (2) contain mainly low frequency intensity distributions with weak color variation and target texture obscured by indirect propagation. Several salient and camouflaged object detection methods respond by producing a compact, rounded mask. Such predictions may locate the dominant light response, but their shapes remain similar for targets with very different structures. RUBIN more frequently recovers the overall height, width, orientation, and occupied area of the target. For elongated examples, it better preserves the principal direction and asymmetric mass distribution; for compact examples, it

Table 6: Quantitative comparison on the NLOS-PASCAL benchmark across seven metrics. The best results are highlighted in **red**, and the second-best results in **blue**. Tied values share the same rank.

Method	Venue	NLOS-PASCAL						
		Max-F↑	Mean-F↑	F_β^w ↑	MAE↓	S_m ↑	E_m ↑	Dice↑
Direct-view segmentation								
LDF	CVPR'20	0.412	0.364	0.286	0.235	0.482	0.628	0.338
ZoomNeXt	TPAMI'24	0.451	0.397	0.319	0.246	0.511	0.624	0.389
GateNet	ECCV'20	0.417	0.369	0.291	0.229	0.486	0.630	0.343
FAP-Net	TIP'22	0.415	0.361	0.279	0.234	0.482	0.622	0.331
PGNet	CVPR'22	0.463	0.422	0.340	0.214	0.538	0.647	0.401
BiRefNet	CAAI AIR'24	0.458	0.411	0.327	0.216	0.532	0.627	0.387
SegFormer	NeurIPS'21	0.458	0.422	0.330	**0.210**	0.539	0.639	0.391
CamoFormer	TPAMI'24	0.469	0.419	0.324	0.213	0.537	**0.653**	0.378
VSCode	CVPR'24	0.451	0.416	0.323	0.221	0.529	0.629	0.384
Indirect-imaging methods								
NLOS-OT	TIP'22	0.404	0.363	0.292	0.242	0.485	0.627	0.352
Learnable Inv. Kernel	ICCV'23	0.424	0.382	0.308	0.229	0.498	0.639	0.365
NeTF	TPAMI'21	0.418	0.377	0.306	0.233	0.495	0.636	0.367
NLOS-LTM	TIP'25	0.444	0.395	0.323	0.237	0.502	0.621	0.385
PAC-Net	CVPR'23	**0.476**	0.427	0.347	0.216	0.536	**0.653**	**0.409**
NLOST	CVPR'23	0.394	0.343	0.265	0.226	0.476	0.620	0.308
LEAP	ECCV'24	0.448	0.400	0.312	0.224	0.514	0.630	0.373
NLOS-LPP	ICCV'25	0.475	**0.428**	**0.348**	0.224	0.538	0.639	0.408
FMoE-NLOS	PR'26	0.474	0.427	0.343	0.215	**0.545**	0.647	0.408
ST-Mamba	NeurIPS'24	0.447	0.386	0.304	0.239	0.500	0.607	0.365
TransiT	ICCV'25	0.450	0.394	0.321	0.233	0.502	0.623	0.381
Ours								
RUBIN	**Ours**	**0.488**	**0.437**	**0.349**	**0.209**	**0.546**	**0.655**	**0.410**

more closely matches the target extent, extending beyond a small central response. This behavior is consistent with the gains in the three F-measures and Dice.

Multiple targets and disconnected components. The advantage becomes more apparent when the ground truth contains multiple instances or disconnected components. Many comparison methods merge nearby targets because indirect propagation mixes their wall responses. Others retain only the strongest component or divide the foreground into unrelated fragments. RUBIN more often preserves the number and approximate placement of the major components. The recovered masks, with coarse boundaries, retain more information about whether the foreground is compact, horizontally

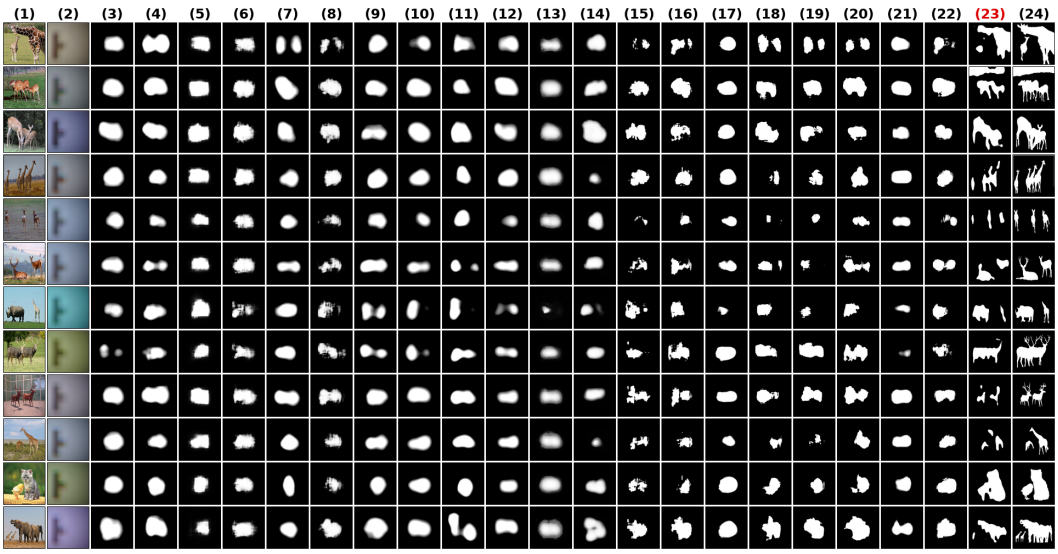

Figure 8: Extended qualitative comparison, Part 1 of 5 (continued).

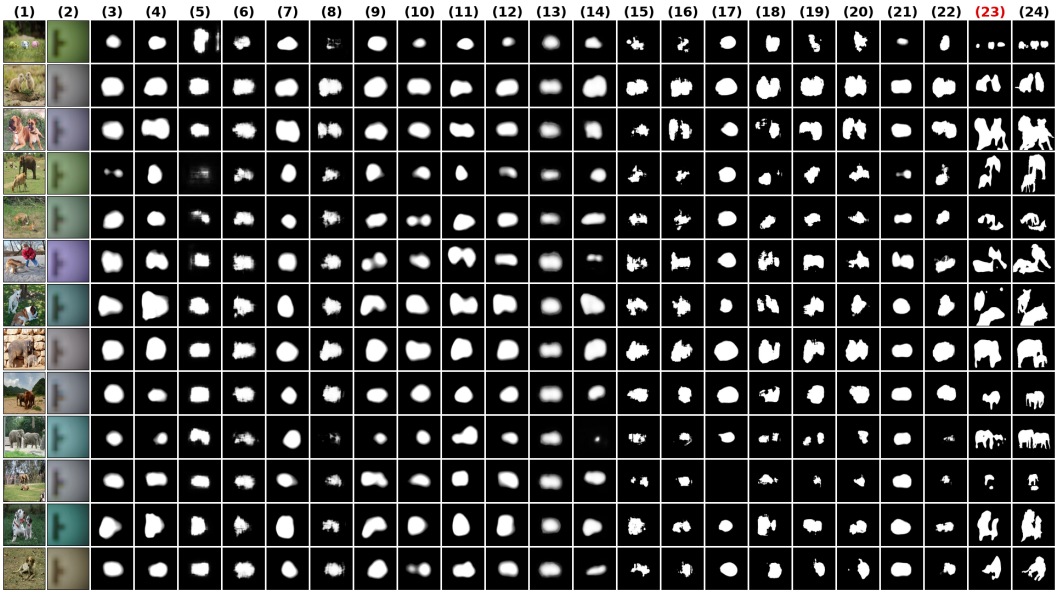

Figure 9: Extended qualitative comparison on measured NLOS observations, Part 2 of 5. The column order and evaluation protocol are identical to those in Figure 8.

distributed, vertically elongated, or separated into several regions. The recovered regions convey the target's basic shape, extent, and spatial arrangement for preliminary inference around corners.

Boundary quality and structural continuity. The comparison reveals a tradeoff between contour smoothness and foreground fragmentation. Some baselines produce stable interiors but remove meaningful protrusions and narrow gaps; others generate sharper masks containing holes, isolated pixels, and disconnected fragments. RUBIN generally retains more continuous interiors together with more target specific boundary variation. This observation is consistent with UGBR, which applies relational correction mainly where the region estimate is uncertain and the boundary branches agree. RCSL provides a complementary explanation for stability across variations in wall brightness, color response, and blur. Together with the component ablations, these observations are consistent with radiance-consistent features and selective boundary and region correction contributing to more coherent foreground recovery.

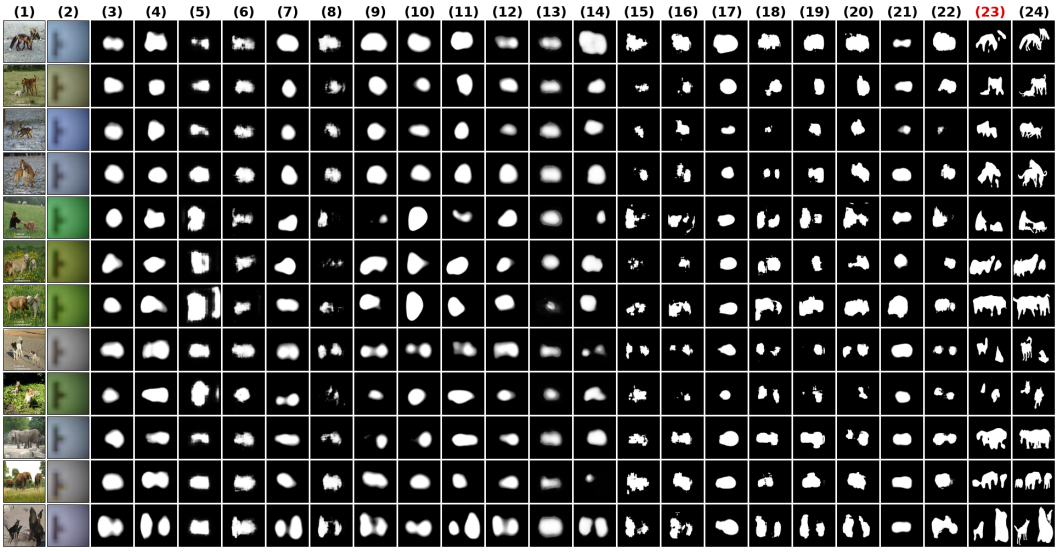

Figure 9: Extended qualitative comparison, Part 2 of 5 (continued).

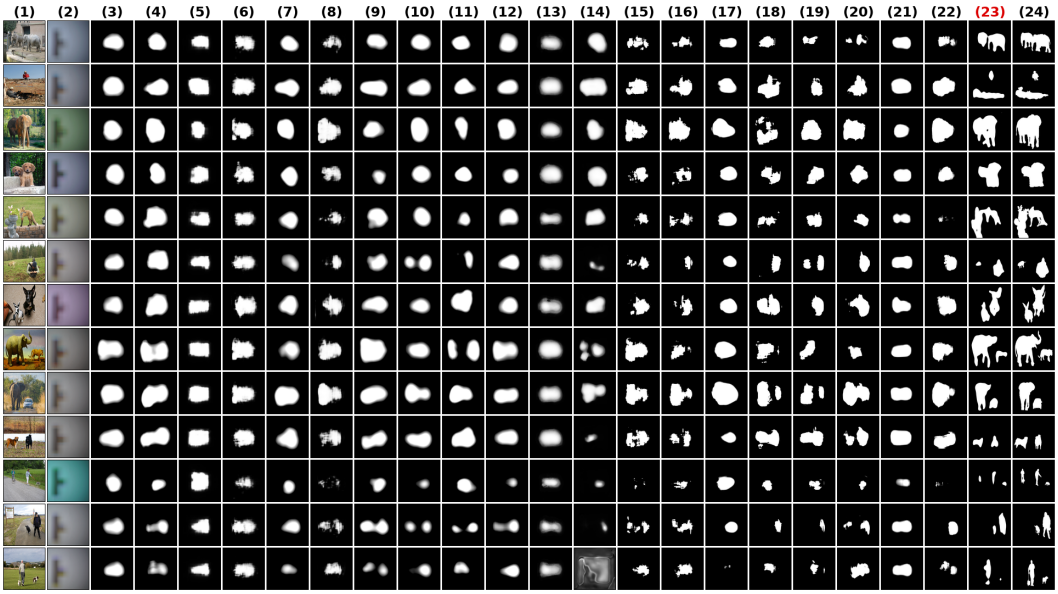

Figure 10: Extended qualitative comparison on measured NLOS observations, Part 3 of 5. The column order and evaluation protocol are identical to those in Figure 8.

Relationship between quantitative and qualitative results. The visual comparisons translate quantitative gains into identifiable shape information. For compact targets, the recovered region more closely matches the occupied area; for elongated targets, it better preserves the principal direction; and for multi-component targets, it more often retains the relative placement of the major regions. These observations connect foreground overlap and structural scores to the basic outline recovered from wall light.

Shape information recovered from indirect wall light. Across the examples, RUBIN recovers the dominant foreground extent, the principal orientation of elongated targets, and the approximate arrangement of major components. These spatial differences make the masks informative about the basic shape behind the corner. Fine protrusions, narrow gaps, and weak components can be smoothed or lost, particularly under weak, saturated, or diffuse observations. The quantitative and

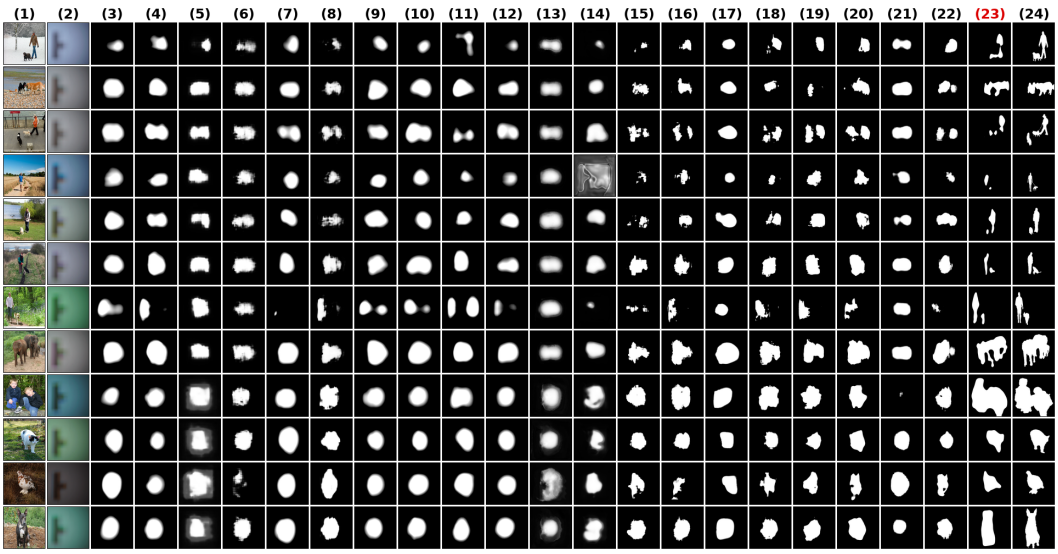

Figure 10: Extended qualitative comparison, Part 3 of 5 (continued).

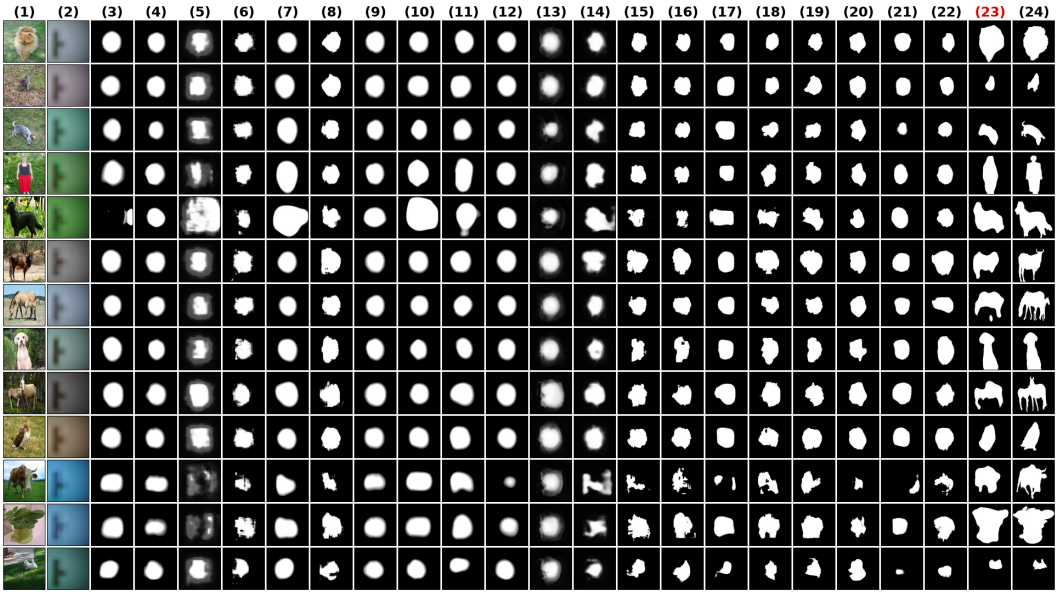

Figure 11: Extended qualitative comparison on measured NLOS observations, Part 4 of 5. The column order and evaluation protocol are identical to those in Figure 8.

visual evidence supports recovery of basic binary silhouettes in the evaluated setup, with foreground extent, contours, and component arrangement as the evaluated shape information.

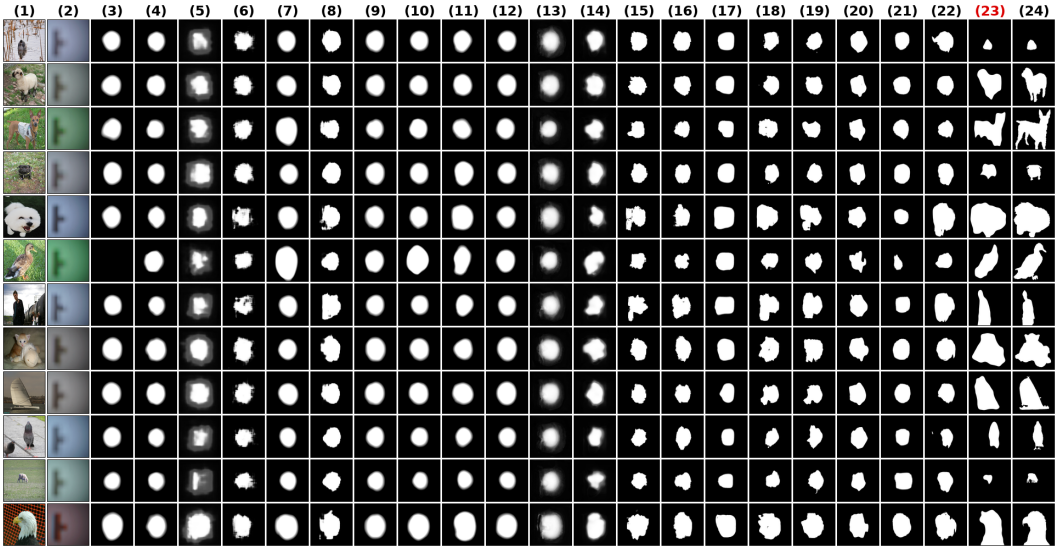

Figure 11: Extended qualitative comparison, Part 4 of 5 (continued).

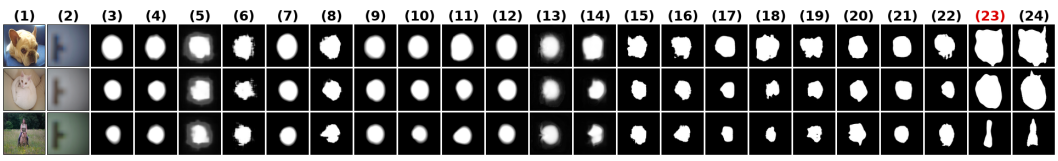

Figure 12: Extended qualitative comparison on measured NLOS observations, Part 5 of 5. The column order and evaluation protocol are identical to those in Figure 8.

E Extended Robustness Analysis

The main paper shows representative predictions under observation variations. This section evaluates computational perturbations of recorded NLOS-ECSSD wall observations, with response curves, summary tables against twenty baselines, and complete qualitative sweeps for one test example. Geometric transformations, radiance changes, and JPEG compression test how foreground recovery responds to changes in the observed wall pattern.

E.1 Evaluation Protocol and Metrics

Controlled perturbations. The evaluation uses all 200 images of the NLOS-ECSSD test split. We apply computational transformations to the recorded RGB wall observations, followed by resizing to 384×384 and evaluation preprocessing. Model parameters remain fixed throughout each sweep. Every method receives the same transformed observation and is evaluated against its corresponding target-coordinate mask. The sweeps examine foreground recovery as the observed pattern changes in geometry, radiance, and compression. Image-space geometry and color transforms approximate selected acquisition effects through computational changes to recorded observations.

Four complementary metrics. We report Max-F, Mean-F, the weighted F-measure F_β^w (labeled Fbw in the existing curves), and MAE. Max-F is the maximum of the curve formed from test-set average precision and recall over thresholds. Mean-F averages per-image F-measures at adaptive thresholds equal to twice the prediction mean, clipped to one. We normalize each prediction and quantize it to 8 bits before computing the metrics. Higher values are better for the three F-measures and lower values are better for MAE. The values in this appendix summarize the computational sweeps; the benchmark reference scores are those in Tables 1 and 4.

Aggregation and ranking. Tables 7, 8, and 9 summarize the same condition-level results shown in the curves. Each score is first computed over the 200 test images at one setting. The left block of each table then takes an unweighted mean over that protocol's settings; the right block takes the minimum for each F-measure and the maximum for MAE. The adverse setting is selected separately for every model and metric, so a row's four worst values may occur under different conditions. For environmental illumination, the mean weights the 25 settings equally. Rankings are determined within each table column, with ties at the reported three-decimal precision sharing a color. The summaries compare average and extreme scores over the specified perturbation settings.

E.2 Camera-Angle Robustness

Camera angle. Pitch, yaw, and roll are swept independently over $-12°, -10°, \ldots, +12°$, giving 13 settings per axis and 39 settings in total. We approximate pitch and yaw with centered shears of coefficient $0.65 \tan(\theta)$ and apply an image-centered rotation for roll. Bilinear interpolation and reflected borders retain the image dimensions. The zero-angle condition occurs once in each axis sweep and is retained in the 39-setting average. These transformations test how foreground recovery responds to changes in the geometry of the observed wall pattern.

Average accuracy and axis dependence. RUBIN retains an accuracy advantage as the wall pattern changes in geometry, reaching mean Max-F of 0.710 compared with the strongest baseline mean of 0.616 (Table 7). Figure 13 separates this response by transformation. RUBIN's mean Max-F is 0.722 for pitch, 0.727 for yaw, and 0.680 for roll; the corresponding mean MAEs are 0.131, 0.128, and 0.151. Roll produces the largest reduction in foreground accuracy among the three image-space transformations.

Accuracy across the angle range. RUBIN's minimum Max-F is 0.616, above the strongest competing minimum of 0.605 (Table 7). Its worst weighted F-measure is 0.448 and its largest MAE is 0.181. The independently selected extrema summarize each metric over the full sweep and may occur at different angles. Figure 13 provides the corresponding response curves.

Structural interpretation. The pitch and yaw sweeps retain useful information about the target's foreground extent under changing wall-image geometry. Roll changes the orientation of the entire

Table 7: Camera-angle robustness on NLOS-ECSSD, aggregating 13 pitch, 13 yaw, and 13 roll settings. Worst values are metric-wise extrema over each model's sweep. Best and second-best results in each column are highlighted in **red** and **blue**. Values tied at three decimal places share the same rank.

Method	Mean over 39 settings				Worst across settings			
	Max-F↑	Mean-F↑	F_β^w ↑	MAE↓	Max-F↑	Mean-F↑	F_β^w ↑	MAE↓
Direct-view segmentation								
LDF	0.605	0.553	0.408	0.241	0.600	0.551	0.405	0.243
ZoomNeXt	0.613	0.552	0.419	0.248	**0.605**	0.545	0.412	0.252
GateNet	**0.616**	0.567	0.423	0.237	0.591	0.547	0.405	0.245
FAP-Net	0.604	0.548	0.406	0.246	0.595	0.542	0.397	0.249
PGNet	0.610	**0.574**	**0.475**	**0.182**	0.575	0.539	0.445	0.197
BiRefNet	0.588	0.542	0.455	0.207	0.580	0.535	**0.446**	0.211
SegFormer	0.592	0.550	0.442	0.211	0.577	0.532	0.421	0.215
CamoFormer	0.562	0.521	0.410	0.206	0.522	0.485	0.378	0.219
VSCode	0.572	0.507	0.391	0.267	0.568	0.504	0.388	0.268
Indirect-imaging methods								
NLOS-OT	0.580	0.546	0.461	0.195	0.543	0.513	0.429	0.212
Learnable Inv. Kernel	0.569	0.538	0.444	0.196	0.515	0.485	0.390	0.219
NeTF	0.567	0.532	0.428	0.195	0.513	0.464	0.347	0.212
NLOS-LTM	0.593	0.557	0.449	0.193	0.570	0.532	0.407	0.210
PAC-Net	0.612	0.572	0.473	0.184	0.588	0.550	0.443	**0.192**
NLOST	0.588	0.551	0.464	0.195	0.550	0.514	0.421	0.213
LEAP	0.602	0.555	0.408	0.245	0.602	0.555	0.407	0.246
NLOS-LPP	0.601	0.557	0.405	0.244	0.598	0.554	0.403	0.245
FMoE-NLOS	0.602	0.552	0.407	0.244	0.602	0.551	0.405	0.245
ST-Mamba	0.601	0.556	0.407	0.243	0.601	0.556	0.406	0.245
TransiT	0.597	0.557	0.406	0.242	0.596	**0.557**	0.405	0.243
Ours								
RUBIN (Ours)	**0.710**	**0.676**	**0.589**	**0.136**	**0.616**	**0.578**	**0.448**	**0.181**

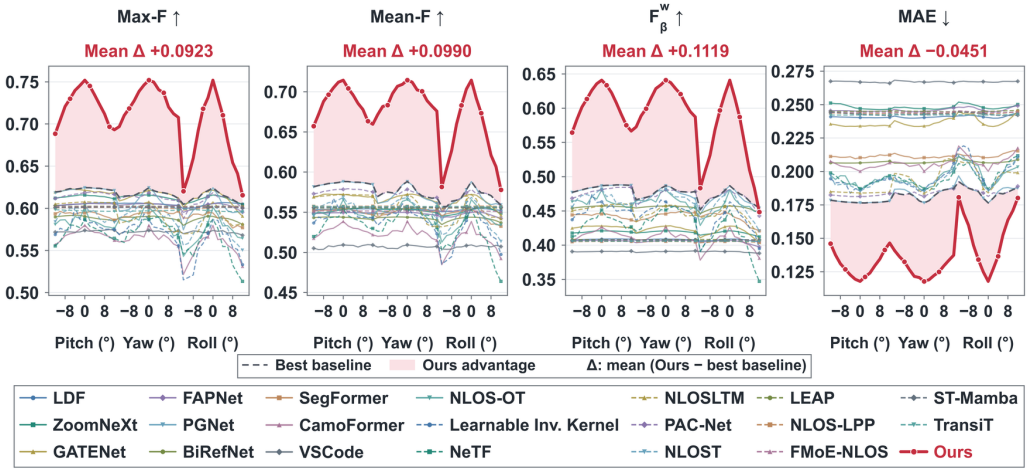

Figure 13: Camera-angle robustness on NLOS-ECSSD. From left to right: Max-F, Mean-F, F_β^w, and MAE. P, Y, and R denote pitch, yaw, and roll; each group runs from $-12°$ to $+12°$. The plot presents three independent axis sweeps consecutively; inter-group segments connect their displayed curves. Every setting is plotted, although only selected ticks are labeled. The red Ours curve denotes RUBIN.

observed pattern, and the corresponding predictions show changes in foreground width and upper extent. The curves and visual sweeps connect geometric variation in the wall observation to the shape recovered behind the corner.

E.3 Environment-Light Robustness

Environmental illumination. The illumination transform modulates the observation with a directional shading field and temperature-dependent RGB gains. Its 25 settings comprise three separate sweeps. Twelve azimuths run from 0° to 330° in 30° steps at unit intensity and 5000 K. The intensity sweep uses six factors, 0.4, 0.6, 0.8, 1.2, 1.4, and 1.6, at a fixed azimuth of 45° and 5000 K. The seven temperatures are 3000, 3500, 4000, 4500, 5500, 6500, and 7500 K, with azimuth and intensity fixed at 45° and one, respectively. A weak directional component remains at unit intensity, so the azimuth sweep still modifies the image. Temperature-dependent RGB gains model changes in color response. Each of the three sweeps varies one parameter while keeping the other two fixed.

Table 8: Environment-light robustness on NLOS-ECSSD. The 25 settings comprise twelve azimuths, six intensities, and seven temperatures. Means weight settings equally; worst values are computed independently for each model and metric. Best and second-best results in each column are highlighted in **red** and **blue**. Values tied at three decimal places share the same rank.

Method	Mean over 25 settings				Worst across settings			
	Max-F↑	Mean-F↑	F_β^w ↑	MAE↓	Max-F↑	Mean-F↑	F_β^w ↑	MAE↓
Direct-view segmentation								
LDF	0.604	0.551	0.405	0.243	0.574	0.524	0.368	0.251
ZoomNeXt	0.613	0.551	0.419	0.248	**0.597**	0.544	0.410	0.257
GateNet	**0.614**	0.565	0.422	0.238	0.546	0.500	0.387	0.264
FAP-Net	0.603	0.545	0.404	0.249	0.571	0.507	0.370	0.279
PGNet	0.609	0.569	0.468	**0.183**	0.504	0.443	0.335	0.220
BiRefNet	0.573	0.524	0.437	0.212	0.426	0.328	0.228	0.241
SegFormer	0.582	0.539	0.432	0.215	0.428	0.384	0.294	0.274
CamoFormer	0.549	0.508	0.395	0.209	0.394	0.341	0.232	0.251
VSCode	0.555	0.490	0.378	0.275	0.327	0.292	0.226	0.330
Indirect-imaging methods								
NLOS-OT	0.585	0.549	0.457	0.191	0.483	0.412	0.287	0.225
Learnable Inv. Kernel	0.585	0.553	0.458	0.188	0.551	0.514	0.406	0.199
NeTF	0.576	0.543	0.442	0.192	0.471	0.439	0.347	0.239
NLOS-LTM	0.603	0.564	0.449	0.187	0.572	0.511	0.369	0.207
PAC-Net	0.613	**0.572**	**0.474**	0.184	0.579	0.524	0.413	**0.198**
NLOST	0.598	0.562	**0.474**	0.188	0.558	0.522	**0.423**	0.205
LEAP	0.601	0.554	0.407	0.245	0.591	0.548	0.403	0.251
NLOS-LPP	0.595	0.552	0.399	0.245	0.502	0.482	0.314	0.253
FMoE-NLOS	0.601	0.551	0.406	0.245	0.592	0.547	0.396	0.251
ST-Mamba	0.601	0.556	0.406	0.244	**0.597**	**0.554**	0.393	0.252
TransiT	0.597	0.556	0.404	0.244	0.594	0.553	0.379	0.255
Ours								
RUBIN (Ours)	**0.732**	**0.695**	**0.617**	**0.127**	**0.652**	**0.609**	**0.521**	**0.168**

Foreground recovery under radiance changes. We test whether foreground recovery persists as the observed radiance changes while the target remains fixed. NLOST uses local and global attention for transient reconstruction (Li et al., 2023a), while ST-Mamba combines neighboring transient frames (Li et al., 2024a). Across the 25 computational settings, RUBIN achieves mean weighted F-measure of 0.617, compared with 0.474 and 0.406 for their RGB-to-mask adaptations, respectively (Table 8). RUBIN leads all four mean-score columns, extending its advantage to the spatial distribution of foreground errors under radiance variation.

Direction and temperature versus intensity. RUBIN's foreground recovery is more stable under direction and color changes than under intensity scaling (Figure 14). Over the twelve azimuths, RUBIN's Max-F ranges from 0.732 to 0.748 and MAE from 0.120 to 0.124. Over the seven temperature settings, these ranges are 0.737–0.747 and 0.122–0.126. In contrast, the intensity subgroup has a mean Max-F of 0.698 and a mean MAE of 0.143. At intensity 1.6, Max-F and Mean-F reach their minima of 0.652 and 0.609, and MAE reaches its maximum of 0.168. The minimum weighted F-measure, 0.521, occurs at intensity 0.4. The subgroup curves distinguish similar responses across direction and color changes from the stronger effect of intensity scaling on foreground extent.

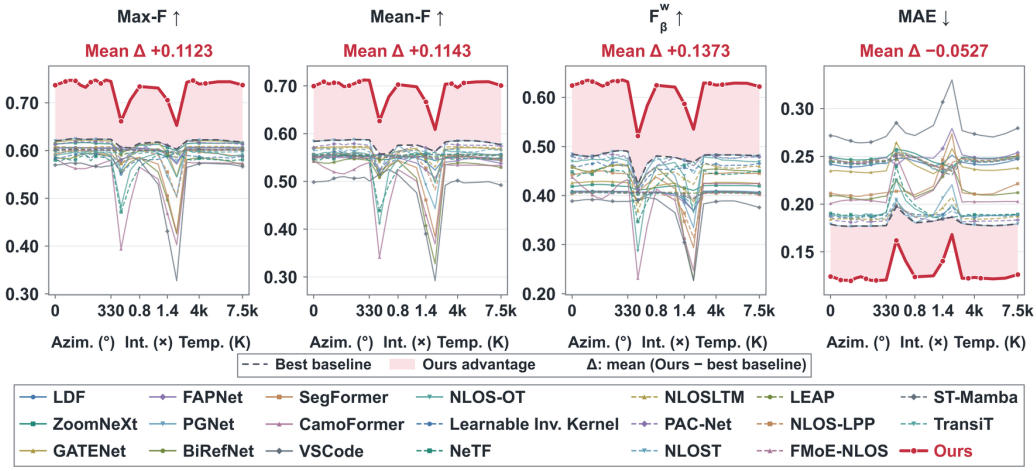

Figure 14: Environment-light robustness on NLOS-ECSSD for Max-F, Mean-F, F_β^w, and MAE. A, I, and K identify azimuth, relative intensity, and temperature. The three independent sweeps are concatenated along their respective azimuth, intensity, and temperature scales. Every setting is plotted, with selected tick labels shown for readability. The red Ours curve denotes RUBIN.

Why the radiance test matters. Directional shading and channel gains change wall-image intensities while the target shape remains fixed. A method that treats these changes as foreground evidence can enlarge a region, suppress weak target support, or introduce holes. RUBIN retains an accuracy advantage in the intensity subgroup, where changes in wall-light amplitudes alter foreground extent and continuity. Together with the RCSL ablations, this result supports an association between radiance-consistent structural features and foreground recovery under changing acquisition radiance.

E.4 JPEG-Compression Robustness

JPEG compression. The sweep comprises ten JPEG qualities, $60, 64, \ldots, 96$, followed by the unchanged reference input, giving eleven settings in total. The ten quality settings apply JPEG encoding and decoding before network preprocessing; the reference is labeled 100 in the existing figures. Lower quality introduces stronger quantization and potential block or ringing artifacts while the target geometry remains fixed. This experiment tests whether weak structural cues remain useful after lossy storage or transmission.

Performance across quality settings. PGNet grafts Transformer features onto a convolutional pyramid (Xie et al., 2022) and achieves the strongest baseline means across all four metrics in this JPEG sweep. Over the eleven settings in Table 9, RUBIN increases mean weighted F-measure from PGNet's 0.487 to 0.605 and lowers mean MAE from 0.177 to 0.131. At quality 60, RUBIN obtains 0.684 Max-F, 0.656 Mean-F, 0.559 weighted F-measure, and 0.148 MAE, which are its reported worst values in the sweep.

Recovered structure as image quality improves. As compression becomes weaker, RUBIN generally recovers a more complete foreground region. Figure 15 shows the accuracy trend, with small local reversals in weighted F-measure and MAE. The visual progression in Figure 20 relates the change to the target's upper contour and foreground extent.

Relation to indirect-light structure. JPEG preserves the overall spatial arrangement while potentially removing low-amplitude variations through quantization. Such variations can matter when the input already contains attenuated target evidence. The wall observations retain useful shape information after encoding and decoding, allowing the recovered silhouette to remain informative across the evaluated quality settings.

Table 9: JPEG-compression robustness on NLOS-ECSSD, over ten JPEG qualities from 60 to 96 in steps of four and one unchanged reference input, labeled 100. Lower quality means stronger compression. Best and second-best results in each column are highlighted in **red** and **blue**. Values tied at three decimal places share the same rank.

Method	Mean over 11 settings				Worst across settings			
	Max-F↑	Mean-F↑	F_β^w ↑	MAE↓	Max-F↑	Mean-F↑	F_β^w ↑	MAE↓
Direct-view segmentation								
LDF	0.606	0.554	0.409	0.240	0.605	0.554	0.409	0.241
ZoomNeXt	0.616	0.553	0.421	0.247	0.615	0.552	0.421	0.247
GateNet	0.621	0.572	0.428	0.234	0.620	0.572	0.427	0.235
FAP-Net	0.606	0.549	0.407	0.245	0.605	0.549	0.407	0.246
PGNet	**0.624**	**0.587**	**0.487**	**0.177**	**0.623**	**0.586**	**0.486**	**0.177**
BiRefNet	0.591	0.544	0.457	0.206	0.590	0.544	0.457	0.207
SegFormer	0.596	0.554	0.446	0.210	0.593	0.551	0.443	0.212
CamoFormer	0.577	0.536	0.424	0.202	0.573	0.532	0.420	0.203
VSCode	0.574	0.507	0.391	0.267	0.572	0.506	0.391	0.267
Indirect-imaging methods								
NLOS-OT	0.580	0.546	0.457	0.195	0.525	0.495	0.399	0.215
Learnable Inv. Kernel	0.585	0.554	0.456	0.187	0.570	0.537	0.437	0.191
NeTF	0.581	0.547	0.448	0.191	0.527	0.497	0.400	0.210
NLOS-LTM	0.608	0.572	0.458	0.184	0.607	0.571	0.457	0.185
PAC-Net	0.618	0.578	0.482	0.182	0.617	0.577	0.481	0.182
NLOST	0.587	0.550	0.458	0.193	0.551	0.516	0.418	0.203
LEAP	0.603	0.555	0.408	0.245	0.602	0.555	0.408	0.245
NLOS-LPP	0.601	0.557	0.405	0.243	0.601	0.557	0.405	0.244
FMoE-NLOS	0.602	0.552	0.407	0.244	0.602	0.551	0.407	0.244
ST-Mamba	0.601	0.556	0.408	0.243	0.601	0.556	0.407	0.243
TransiT	0.597	0.557	0.406	0.242	0.597	0.557	0.406	0.243
Ours								
RUBIN (Ours)	**0.722**	**0.688**	**0.605**	**0.131**	**0.684**	**0.656**	**0.559**	**0.148**

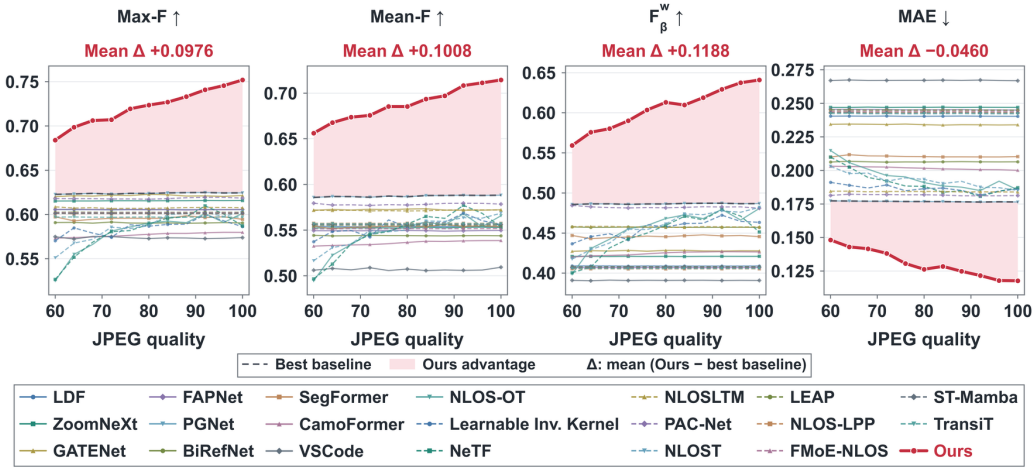

Figure 15: JPEG-compression robustness on NLOS-ECSSD for Max-F, Mean-F, F_β^w, and MAE. Quality increases from 60 on the left to 100 on the right, so compression becomes weaker toward the right. Quality 100 is the unmodified reference observation. The red Ours curve denotes RUBIN.

E.5 All-Setting Qualitative Comparisons

Figures 16, 17, 18, 19, and 20 show the complete sweeps for NLOS-ECSSD sample 0022. Each row changes the acquisition perturbation while keeping the target fixed. The 23-column layout differs from the 24-column comparison in Appendix D.2: it starts directly with the perturbed relay-wall

observation. Columns (2)–(21) contain the twenty baselines in the table order, column (22), marked in red, is RUBIN, and column (23) is the unchanged target mask. The column mapping is specified in Figure 16. These plates trace shape changes within the same example, while the 200-image curves and tables quantify dataset-level recovery.

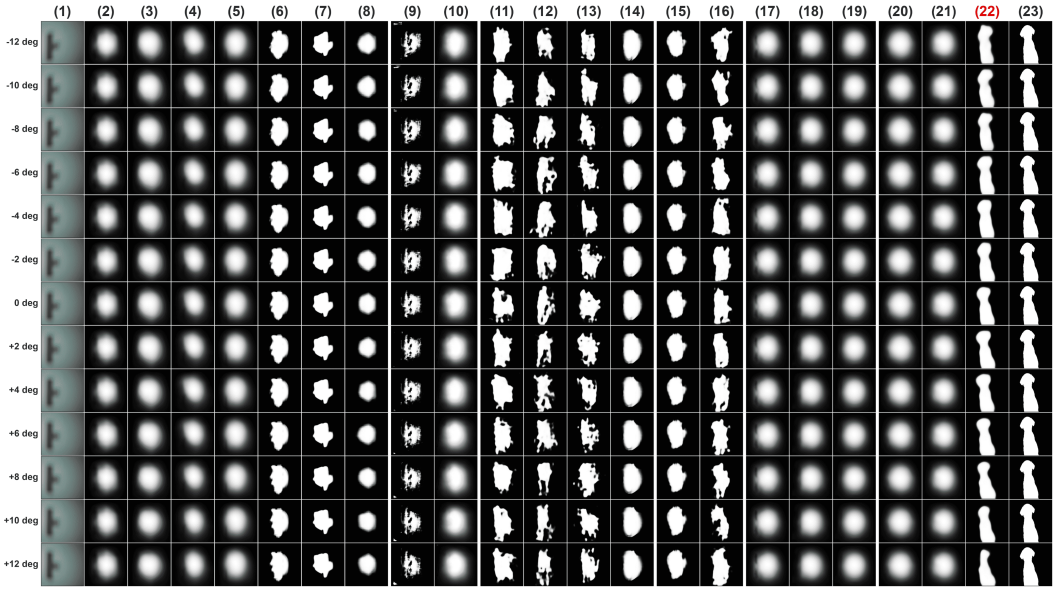

Figure 16: Complete pitch sweep for NLOS-ECSSD sample 0022, from -12° to $+12^\circ$ in 2° steps. Columns are (1) perturbed wall observation; (2)–(21) LDF, ZoomNeXt, GateNet, FAP-Net, PGNet, BiRefNet, SegFormer, CamoFormer, VSCode, NLOS-OT, Learnable Inverse Kernel, NeTF, NLOS-LTM, PAC-Net, NLOST, LEAP, NLOS-LPP, FMoE-NLOS, ST-Mamba, and TransiT; (22) RUBIN; (23) target ground truth. All methods receive only column (1). This order is shared by the following robustness plates.

Pitch and yaw: stable foreground extent. Across Figures 16 and 17, RUBIN retains the dominant elongated foreground and its approximate vertical extent. Several baselines maintain a broad rounded response, while others produce sharper but irregular or perforated regions. The recovered RUBIN region continues to convey the target's main shape and orientation, with fine boundary details remaining smoother than the target mask.

The yaw sweep in Figure 17 varies the other shear direction while keeping the target fixed. Together with Figure 13, the pitch and yaw sweeps show how useful foreground structure persists as the observed wall-image geometry changes.

Roll: changes in recovered shape. The roll sweep in Figure 18 connects changes in wall-pattern orientation to changes in the recovered foreground. RUBIN captures a coherent elongated region around the central orientations. Large negative rotations broaden or separate the upper response, while large positive rotations narrow the prediction and reduce its upper extent. These observations link the lower roll averages to geometric misalignment of the wall pattern.

Foreground recovery across illumination changes. In Figure 19, the recovered silhouettes remain similar across azimuth and temperature changes, while intensity scaling changes their width and upper extent. Some comparison methods develop broken regions or stronger foreground shrinkage at the intensity extremes. RUBIN continues to recover a dominant foreground region, connecting the subgroup curves to visible information about target shape.

Foreground continuity under compression. RUBIN retains a connected elongated foreground throughout the quality sweep in Figure 20. Weaker compression produces a more complete upper contour and greater foreground extent, consistent with the accuracy trend in Figure 15. The persistent main component provides useful shape information as image quality changes.

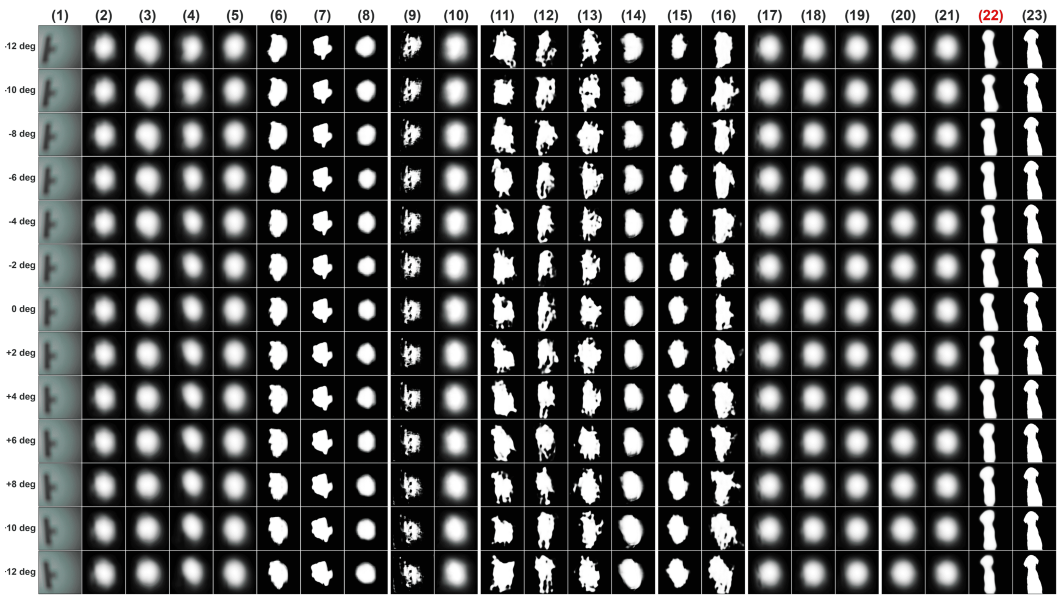

Figure 17: Complete yaw sweep for NLOS-ECSSD sample 0022, from -12° to $+12^\circ$ in 2° steps. The column order follows Figure 16. RUBIN retains the main elongated support across this example's sweep, with a smoother contour than the target mask.

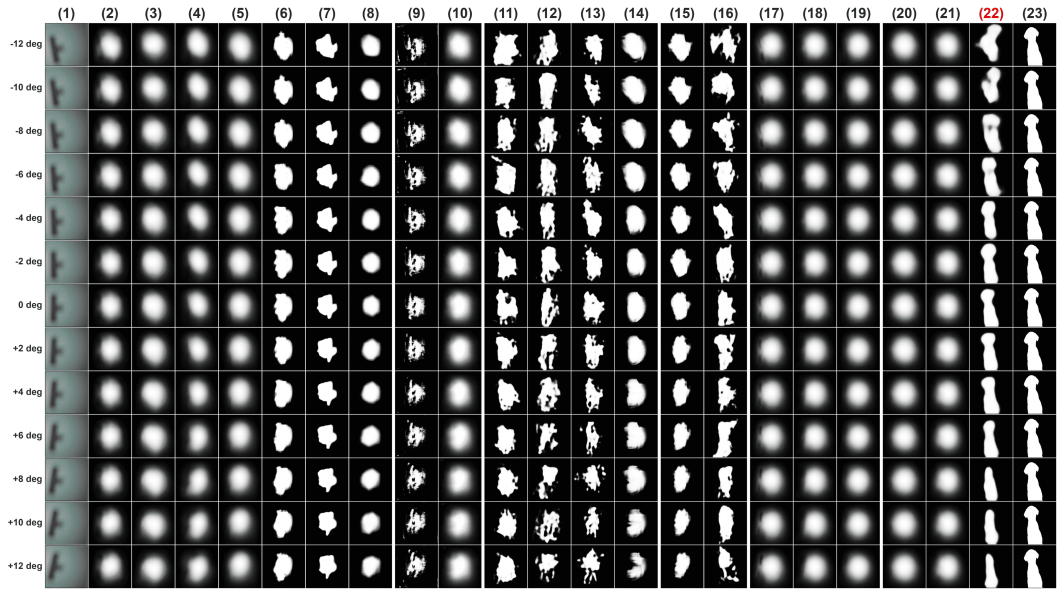

Figure 18: Complete roll sweep for NLOS-ECSSD sample 0022, from -12° to $+12^\circ$ in 2° steps. Columns follow Figure 16. Extreme negative roll distorts or splits RUBIN's response, while extreme positive roll suppresses the upper target extent. The full sweep extends the near-zero example with angle-dependent changes in recovered extent and continuity.

E.6 Structural Recovery across Observation Changes

Across geometric, radiance, and JPEG perturbations, RUBIN ranks first in all twelve mean-score columns and all twelve worst-setting columns of Tables 7–9. The 200-image evaluation and complete visual sweeps support the persistence of useful foreground structure as the observed wall pattern changes. Together with the component ablations, these results support the complementary roles of radiance-consistent structural learning and uncertainty-gated refinement in recovering basic hidden-target shape as wall observations change.

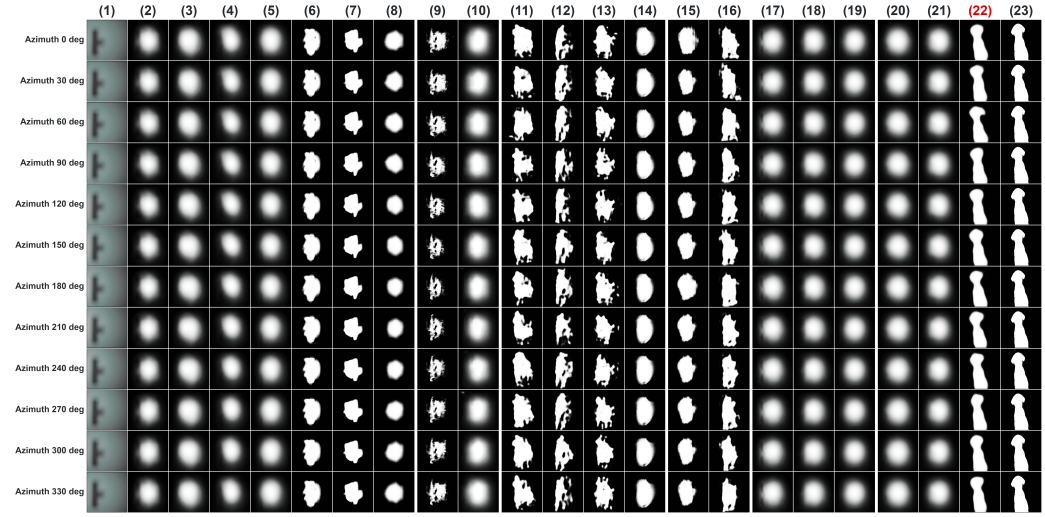

Figure 19: Complete environment-light sweep for NLOS-ECSSD sample 0022. Across this panel and its continuation, rows show twelve azimuths, six relative intensities, and seven temperatures in that order. Columns follow Figure 16; the target mask remains unchanged. RUBIN's largest support changes occur in the intensity subgroup, while many baseline masks also develop fragmentation or strong foreground shrinkage.

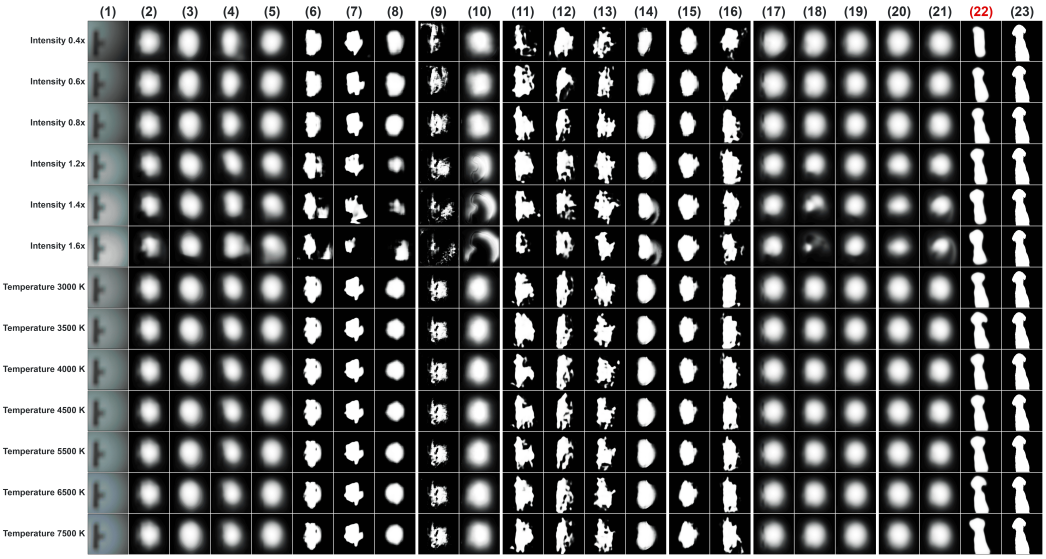

Figure 19: Environment-light sweep (continued): intensity and temperature.

The quantitative evaluation concerns three independently applied computational perturbation families on one NLOS-ECSSD split. The qualitative sweeps use one example, and the reported runs do not estimate uncertainty across training seeds. Large rotations and strong compression reduce recovered foreground extent. Within these evaluated conditions, the recovered foreground structure supports the feasibility of inferring basic hidden-target shape from ordinary-camera wall observations.

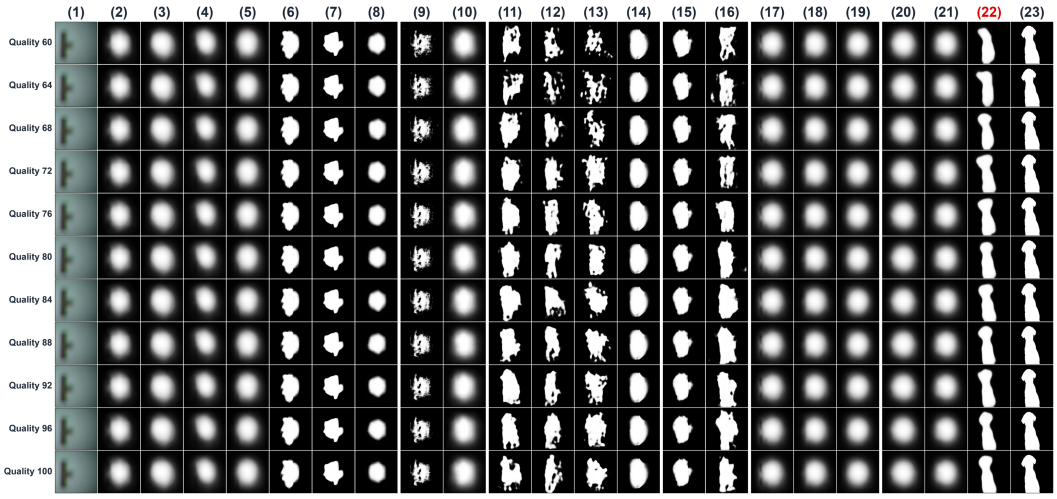

Figure 20: Complete JPEG sweep for NLOS-ECSSD sample 0022 at qualities 60, 64, . . ., 100. Columns follow Figure 16. RUBIN's main component persists across the quality sweep; stronger compression changes its upper contour and reduces its recovered extent. Quality 100 is the unchanged observation.