photonics

Article

Bounded Hidden-Screen Reconstruction from Diffuse Wall Reflections via Survivability-Guided Representation Learning

Jie Zhou [1], Zhiwen Zheng [2], Wenwen Tang [3], Xingru Huang [2] and Huiyu Zhou [4,*]

[1] Faculty of Engineering, The University of Sydney, Sydney, NSW 2006, Australia; stellazhou116@gmail.com
[2] School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China; zhiwen.zheng@hdu.edu.cn (Z.Z.); xingru.huang@hdu.edu.cn (X.H.)
[3] School of Education, Johns Hopkins University, Baltimore, MD 21218, USA; wtang18@jhu.edu
[4] School of Computing and Mathematical Sciences, University of Leicester, Leicester LE1 7RH, UK
* Correspondence: hz143@leicester.ac.uk

Abstract

Reconstructing a hidden-screen image from one ordinary Red-Green-Blue (RGB) photograph of a diffuse wall when the screen is outside the camera's direct line of sight is a challenging problem in passive indirect imaging that highlights potential visual-information leakage. This task presents two main challenges: diffuse reflection unevenly attenuates spatial cues, and different screen content may require observation-dependent rather than fixed feature aggregation. To address these issues, we introduce Survivability-Guided Representation Learning (SGRL). Our method features Survivability-Band Representation (SBR), which uses two Gaussian-initialized trainable depthwise operators to form three nominal complementary construction paths, LOW, MIDDLE, and HIGH. Content-Conditioned Cross-Band Fusion (CCF) then predicts observation-conditioned computational coefficients to combine the branch features and generate the hidden-screen reconstruction. Experiments under controlled passive imaging conditions show that SGRL achieves the strongest recorded reconstruction performance among the evaluated controls across multiple metrics, supporting the structure-level recovery of broad brightness distribution, region layout, and major visual structure from diffuse wall observations.

Keywords: survivability-guided representation learning; frequency-inspired representation; diffuse optical transport; hidden-screen reconstruction; image restoration

1. Introduction

In real-world environments such as open offices or shared spaces, the displayed content of a target screen may be outside the camera's direct view, while its emitted light is projected and scattered onto nearby walls, as shown in Figure 1a. An observer may photograph only the wall with an ordinary Red-Green-Blue (RGB) camera and still record faint brightness and color variations induced by the hidden screen, creating a potential visual-information leakage channel. This setting belongs to the broader class of passive Non-Line-of-Sight (NLOS) imaging problems, in which hidden scene information is inferred from indirect environmental reflections. Accordingly, this paper studies whether the broad visual structure of a hidden screen can be recovered from a single ordinary RGB photograph of the diffuse wall.

As illustrated in Figure 1b,c, the acquisition relies only on the indirect screen–wall–camera light path, and the camera observes a blurred wall response rather than the screen itself. Under the evaluated controlled conditions, with no active illumination, video sequence, or precise

Check for updates

Received: 27 August 2026
Revised: 10 September 2026
Accepted: 11 September 2026
Published: 21 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

geometric calibration, we focus on the structure-level recovery of brightness distribution, region layout, and major visual contours.

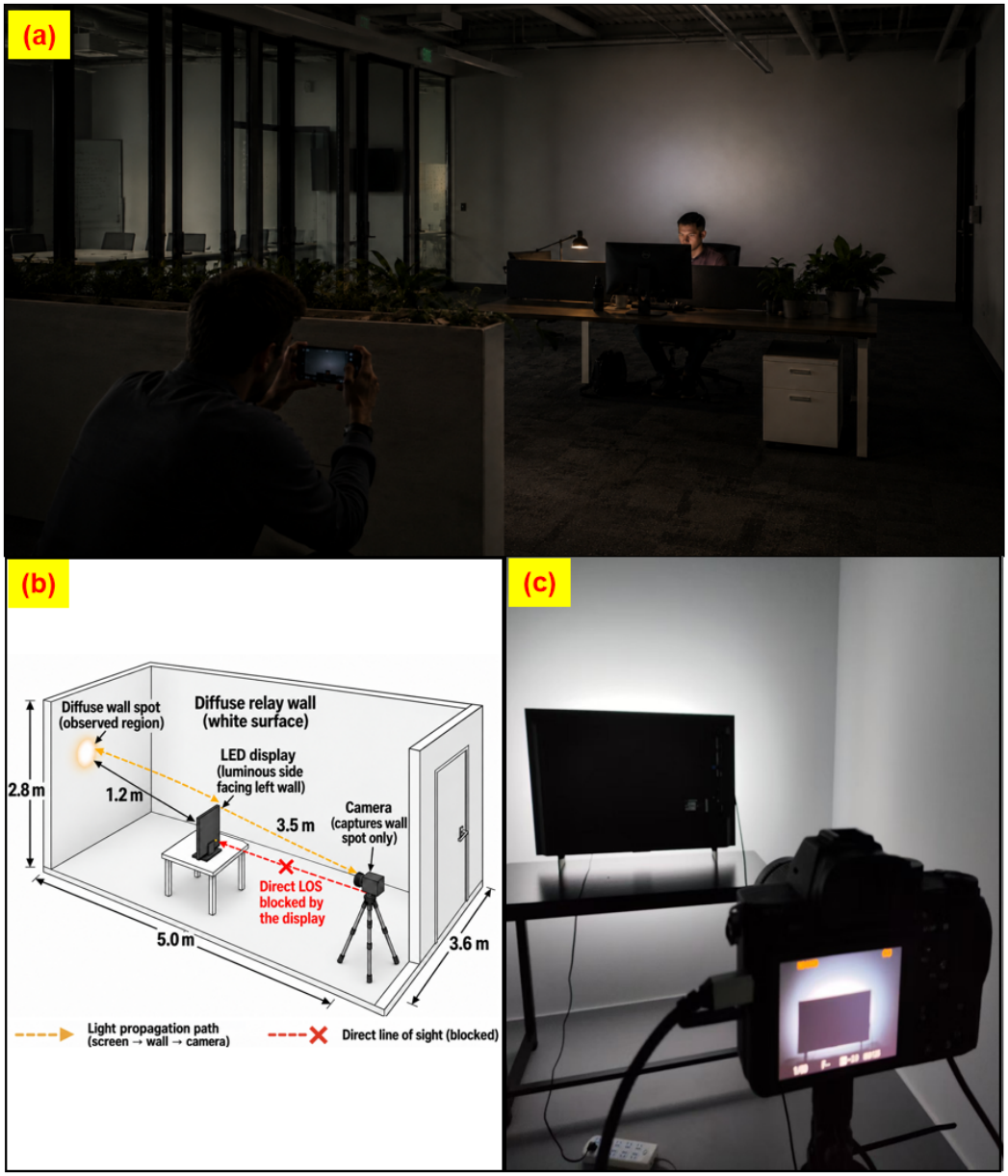


Figure 1. Passive hidden-screen reconstruction scenario and experimental setup. (**a**) Representative privacy scenario in which a screen outside the camera's direct view leaves indirect optical traces on a nearby wall. (**b**) Controlled acquisition geometry showing the indirect screen–wall–camera propagation path and the blocked direct line of sight. (**c**) Physical acquisition setup and the corresponding diffuse-wall observation recorded by the Sony A7S II RGB camera (Sony Corporation, Tokyo, Japan).

Reconstructing hidden scenes from faint diffuse reflections introduces two computational challenges. First, scattering and spatial mixing attenuate spatial cues unevenly, with the extent of preservation depending on the scene configuration and the content itself. Second, the useful evidence may differ across screen content, so a fixed feature-aggregation rule may not respond equally well to all observations. These challenges motivate an architecture that separates complementary representations while conditioning their computational combination on the current wall observation.

To address these challenges within the evaluated setting, we propose Survivability-Guided Representation Learning (SGRL). Its Survivability-Band Representation (SBR) uses

two Gaussian-initialized trainable depthwise operators to form three nominal complementary construction paths, LOW, MIDDLE, and HIGH. Content-Conditioned Cross-Band Fusion (CCF) then predicts observation-conditioned computational coefficients to combine the branch features. The branches are architectural representations rather than verified physical passbands, and the CCF coefficients are not calibrated optical reliability or signal-to-noise estimates. The resulting contribution is a bounded learned wall-to-screen inverse for structure-level reconstruction under controlled passive imaging conditions.

The main contributions are summarized as follows:

(1) We study bounded single-frame passive RGB hidden-screen reconstruction from one ordinary diffuse-wall photograph under controlled no-direct-view conditions, and characterize the associated visual-information leakage without claiming unrestricted recovery.
(2) We propose the SBR, which organizes the wall observation into three nominal complementary construction paths using Gaussian-initialized trainable depthwise operators.
(3) We propose the CCF module, which uses observation-conditioned computational coefficients to combine the SBR branch features for the tested wall-to-screen reconstruction task.

2. Related Work

Three areas provide the closest context for this work: passive NLOS imaging, optical side-channels, and learning-based image restoration. They respectively establish that hidden scenes can leave measurable evidence on relay surfaces, that optical emissions may unintentionally carry information, and that learned priors can recover images from severe degradation. Our task intersects all three directions but seeks a dense RGB screen estimate from only one diffuse-wall observation.

2.1. Active and Passive NLOS Imaging

Conventional active NLOS imaging relies on controlled illumination and often employs time-resolved sensing to recover hidden geometry or reflectance [1,2]. These measurements can provide informative transport cues, but they generally require specialized illumination, sensing hardware, and acquisition or calibration procedures that are absent from the passive setting considered here.

Passive NLOS imaging instead uses naturally available ambient or existing light together with spatial or temporal evidence. Examples include accidental pinholes and pinspecks, visible corners, relay-surface observations, and blank-wall evidence [3–6]. These approaches make different geometric and scene assumptions and support different outputs; importantly, not all passive methods require temporal cues or strict calibration.

We position this paper as a bounded passive NLOS inverse problem that maps one ordinary RGB photograph of one diffuse relay wall to a dense paired RGB hidden-screen reconstruction with the screen outside the camera's direct view. The method uses no active illumination, transient sensing, video sequences, calibrated light transport, or precise geometry. Thus, the task is a cross-domain wall-to-screen inverse rather than conventional same-domain image restoration, and the claim is limited to structure-level recovery under the evaluated controlled conditions rather than a universal solution for passive NLOS imaging.

2.2. Optical Side-Channels

Optical side-channel studies show that emitted, reflected, or modulated light can carry unintended information. Device indicators and Cathode-Ray Tube (CRT) displays have been shown to reveal data-dependent optical signals [7,8], while compromising reflections can preserve a recognizable display structure [9]. Lamphone demonstrates another form of

optical leakage in which reflected light from a vibrating bulb carries recoverable acoustic information [10].

The diffuse-wall channel considered here is more spatially lossy than a specular reflection. Wall scattering mixes light over space and direction, so the observed image cannot be treated as a geometrically warped screen image. Broad color and layout variations may leave more detectable wall variation than text strokes and thin boundaries, although that ordering is not measured here as a physical transfer function. We therefore study bounded reconstruction from the recorded wall evidence rather than assuming unrestricted recovery of arbitrary screen content.

2.3. Learning-Based Image Restoration

Learning-based restoration provides strong priors for recovering images from blur, noise, resolution loss, and mixed degradations. Transformer-based methods such as Restormer [11], SwinIR [12], and Uformer [13] model local and nonlocal image dependencies, while AirNet [14] and PromptIR [15] address unknown or mixed degradation through learned conditioning. Our experimental controls further include DRCT [16], NGswin [17], OSRT [18], SFHformer [19], BGAI [20], AdaRevD [21], DnCNN [22], DAN variants [23], SPAN [24], CFSR [25], DiffBIR [26], OSEDiff [27], ResDiff [28], SinSR [29], DeepSN-Net [30], and Pixel2Pixel [31].

These methods provide an architecture-diverse reconstruction context, but their inclusion does not isolate capacity or training-recipe effects. Conventional restoration normally assumes that the degraded input and clean target remain in approximately the same image domain. Here, the input belongs to the relay-wall domain and the target belongs to the hidden-screen domain, with an unknown optical operator mixing screen radiance before capture. SGRL addresses this mismatch by organizing measured wall evidence with a frequency-inspired complementary representation before cross-band synthesis.

3. Method

3.1. Physical Image Formation and Inverse Objective

Let $L(u, \lambda)$ be the spectral radiance emitted at screen coordinate u, and let $X_c(p)$ be the value recorded at camera pixel p in color channel c. Screen light first propagates to relay-wall point r, is modulated by visibility and soft edge-transition effects associated with the finite source and occluder geometry, is redistributed by the wall bidirectional reflectance, and is finally integrated by the camera optics and spectral response. Under incoherent illumination, this path is described by

$$\begin{aligned} J(r,\lambda) &= \int_{\Omega_s} V(u,r) P_{\mathrm{pen}}(u,r) G(u,r) L(u,\lambda)\, du, \\ X_c(p) &= \int_{\Lambda} q_c(\lambda) \int_{\Omega_w} h_c(p,r,\lambda) \frac{\rho_w(r,\lambda)}{\pi} \\ &\quad \times J(r,\lambda)\, dr\, d\lambda + b_c(p) + n_c(p). \end{aligned} \tag{1}$$

Here, J is the irradiance incident on the relay wall, q_c is the camera spectral sensitivity, h_c is the wall-to-sensor point-spread function, and ρ_w/π is the Lambertian component of wall reflectance. V denotes binary or fractional visibility, whereas P_{pen} is a phenomenological factor for soft penumbral and edge-transition modulation induced by the finite source and occluder geometry. G contains inverse-square and cosine falloff, and b_c and n_c denote slowly varying photometric drift and sensor noise. We do not explicitly model or claim a measurable wave-optical edge-diffraction contribution in the present acquisition system. This continuous formulation identifies the information bottleneck; it is not a fully calibrated room model. The composite discretized forward operator is reserved for $\mathcal{A}_\zeta$ below.

The factors in this model play different roles. Visibility, occlusion, and display-edge geometry introduce spatially structured modulation before the signal reaches the relay wall. Wall reflectance and multipath transport then spread contributions from different screen regions, while the camera optics, spectral response, exposure, and sensor noise further integrate and perturb the result. The combined operator is therefore spatially varying and generally unknown. Equation-level modeling is used here to identify the information bottleneck and to motivate the representation design; SGRL does not require explicit calibration of every physical parameter at test time.

After color integration and discretization, the forward process can be written compactly as

$$x = \mathcal{A}_\zeta y + b + n, \qquad n \sim \mathcal{N}(0, \Sigma_n), \tag{2}$$

where y is the hidden RGB screen, x is the observed wall image, $\mathcal{A}_\zeta$ is the composite light-transport operator, and ζ collects unknown geometry, visibility, wall, and camera parameters. If $\mathcal{A}_\zeta = U\,\mathrm{diag}(\sigma_i)V^\mathsf{T}$, direct inversion amplifies noise in mode i by σ_i^{-1}. Diffuse transport drives many σ_i toward zero, so an unconstrained inverse has variance proportional to $\sum_i \sigma_i^{-2}$ and is physically unstable.

This instability has two consequences. Noise amplification is one: small perturbations in weak measurement modes can produce large changes in an unconstrained reconstruction. Non-identifiability is the other: when distinct screen patterns induce nearly identical wall measurements, no estimator can determine their exact high-frequency difference from the observation alone. The role of learning is therefore not to eliminate the physical information boundary but to use paired data to regularize the inverse toward screen structures that remain compatible with the measured evidence.

The same limitation appears in the noise-weighted information operator

$$\begin{aligned} \mathcal{I}_\zeta &= \mathcal{A}_\zeta^* \Sigma_n^{-1} \mathcal{A}_\zeta, \\ \mathfrak{q}(v) &= \frac{\langle v, \mathcal{I}_\zeta v \rangle}{\langle v, v \rangle}, \end{aligned} \tag{3}$$

where $\mathfrak{q}(v)$ measures the recoverable information along screen mode v. Modes with a small $\mathfrak{q}$ cannot be stably reconstructed from the wall measurement, regardless of estimator capacity. SGRL therefore targets a regularized maximum a posteriori estimate rather than an exact inverse:

$$\begin{aligned} \widehat{y} = \arg\min_z \frac{1}{2} \left\| \Sigma_n^{-1/2} (\mathcal{A}_\zeta z + b - x) \right\|_2^2 \\ + \mathcal{J}_{\mathrm{surv}}(z \mid x). \end{aligned} \tag{4}$$

The data term enforces compatibility with the observed wall image, whereas $\mathcal{J}_{\mathrm{surv}}$ denotes a learned, observation-conditioned regularizer. SGRL realizes this regularization through a trainable complementary representation and observation-conditioned synthesis rather than through one fixed smoothness penalty. SBR organizes the available evidence without discarding the exact complementary sum, and CCF adapts the computational contribution of the branch features. This formulation is a regularized inverse objective motivated by the information bottleneck; it does not claim exact recovery of unobserved screen details or an exact calibrated physical inverse.

The physical model motivates the representation organization. The actual trainable realization of SBR and CCF is specified below using the terminology and evidence boundaries of the present manuscript.

Figure 2 provides a unified view of the physical motivation, SBR, CCF, and bounded output. Its frequency- and reliability-related annotations are interpreted only as design abstractions under the calibration limits stated in the caption and the method definitions.

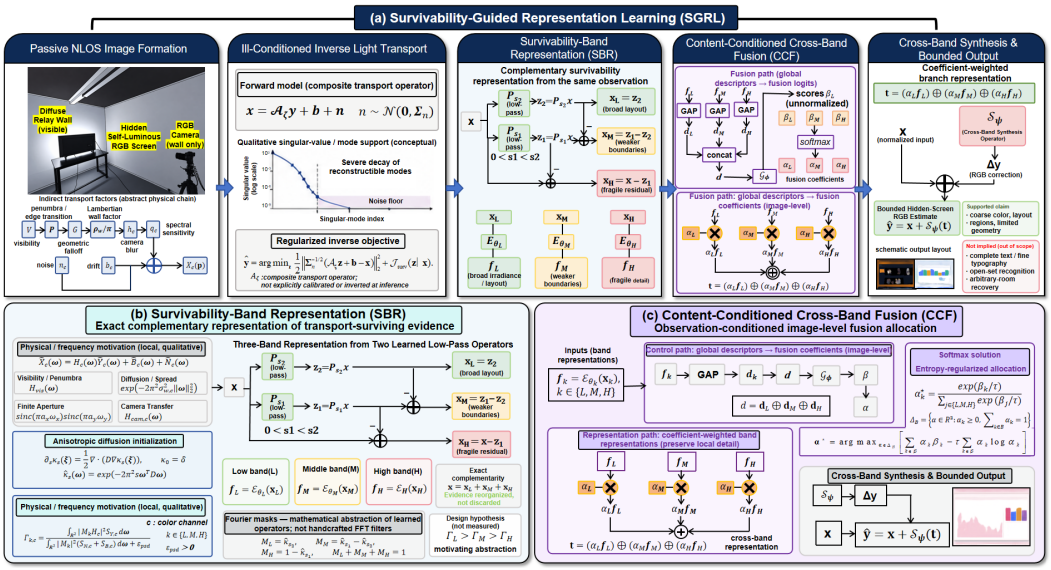


Figure 2. Overview of Survivability-Guided Representation Learning. Panel (**a**) illustrates the passive image-formation model and regularized inverse formulation. Panel (**b**) shows the Survivability-Band Representation, where LOW, MIDDLE, and HIGH denote nominal complementary branches and the illustrated $\Gamma_L > \Gamma_M > \Gamma_H$ relation is a design hypothesis rather than a measured physical ordering. Panel (**c**) shows Content-Conditioned Cross-Band Fusion, whose fusion weights are observation-conditioned computational coefficients rather than calibrated optical reliability or Signal-to-Noise Ratio (SNR). Arrows indicate computational flow, while the colors distinguish the LOW, MIDDLE, HIGH, and CCF pathways. The superscript * in α^* denotes the optimizer of the corresponding allocation problem.

3.2. Survivability-Band Representation

In the process of reconstructing a hidden-screen image from a wall observation, SBR organizes visual cues that undergo different degrees of attenuation in the diffuse-wall image. Specifically, it uses two trainable depthwise convolution operators initialized with Gaussian kernels to construct three exactly complementary representation branches, LOW, MIDDLE, and HIGH, and learns features for each branch separately. The three branches are used to represent large-scale color and layout, transitional structures, and local edges and details, respectively. In this way, residual cues at different scales can be processed along corresponding representation paths, providing structured inputs for subsequent feature fusion.

SBR is a frequency-inspired complementary representation; frequency decomposition is its construction mechanism rather than a measured physical result. Its construction is motivated by a local low-pass approximation of the passive path. Under a locally shift-invariant abstraction,

$$\widehat{X}_c(\omega) = H_c(\omega)\widehat{Y}_c(\omega) + \widehat{B}_c(\omega) + \widehat{N}_c(\omega), \tag{5}$$

where $\omega = (\omega_x, \omega_y)$ is spatial frequency and H_c denotes an unmeasured composite transfer function in this abstraction. A motivating approximation separates several factors:

$$\begin{aligned} |H_c(\omega)| \approx{}& |H_{\mathrm{vis}}(\omega)| \exp\!\left(-2\pi^2\sigma_{w,c}^2\|\omega\|_2^2\right) \\ &\times \left|\mathrm{sinc}(\pi a_x\omega_x)\,\mathrm{sinc}(\pi a_y\omega_y)\right| |H_{\mathrm{cam},c}(\omega)|. \end{aligned} \tag{6}$$

The visibility and penumbra term H_{vis}, diffuse-wall spread $\sigma_{w,c}$, finite sensor aperture (a_x, a_y), and camera transfer $H_{\text{cam},c}$ illustrate why broad color and layout modes may be more detectable than narrow strokes. This local approximation is a design motivation, not a calibrated transfer model of the room: the true path is not assumed to be Gaussian or globally shift-invariant, and its spatial-frequency response was not directly measured.

Let $\kappa_s(\xi)$ denote an anisotropic diffusion kernel at scale s, where ξ is the kernel displacement and $D \succeq 0$ is the motivating anisotropic diffusion tensor:

$$\begin{aligned} \partial_s \kappa_s(\xi) &= \frac{1}{2} \nabla \cdot (D \nabla \kappa_s(\xi)), \qquad \kappa_0 = \delta, \\ \widehat{\kappa}_s(\omega) &= \exp\Big(-2\pi^2 s \omega^{\mathsf{T}} D \omega\Big). \end{aligned} \tag{7}$$

SGRL uses two ordered scales $0 < s_1 < s_2$ to initialize discrete trainable depthwise low-pass operators $\mathcal{P}_{s_1}$ and $\mathcal{P}_{s_2}$. Their learnable parameters are η_1 and η_2, respectively. The 7×7 and 15×15 initial kernels use $\sigma = 1.5$ and $\sigma = 4.0$, respectively, and remain trainable after initialization.

The input observation is decomposed into three nominal construction paths:

$$\begin{aligned} x_L &= \mathcal{P}_{s_2} x, \\ x_M &= (\mathcal{P}_{s_1} - \mathcal{P}_{s_2}) x, \\ x_H &= (I - \mathcal{P}_{s_1}) x. \end{aligned} \tag{8}$$

For the ordered Gaussian initialization, the corresponding masks are

$$\begin{aligned} M_L &= \widehat{\kappa}_{s_2}, \qquad M_M = \widehat{\kappa}_{s_1} - \widehat{\kappa}_{s_2}, \\ M_H &= 1 - \widehat{\kappa}_{s_1}, \qquad M_L + M_M + M_H = 1. \end{aligned} \tag{9}$$

Thus, $x = x_L + x_M + x_H$ exactly. LOW, MIDDLE, and HIGH are nominal construction paths. Because the operators remain trainable, the present study does not claim that their post-training responses are non-negative, strictly ordered, or verified physical passbands.

Exact complementarity is important because the decomposition is used as evidence organization rather than early evidence rejection. Overlapping branches would repeatedly present the same wall energy to multiple estimators, whereas non-complementary masks could remove weak but potentially useful residue before adaptive fusion. This exact complementarity ensures that every measured component remains available to the model; the subsequent estimators decide how that component should be regularized.

For design interpretation only, a hypothetical transported signal-to-corruption ratio in nominal branch k can be written as

$$\Gamma_{k,c} = \frac{\int_{\mathbb{R}^2} |M_k H_c|^2 S_{Y,c} \, d\omega}{\int_{\mathbb{R}^2} |M_k|^2 (S_{N,c} + S_{B,c}) \, d\omega + \varepsilon_{\text{psd}}}. \tag{10}$$

Here, $S_{Y,c}$, $S_{N,c}$, and $S_{B,c}$ denote screen-signal, sensor-noise, and slowly varying background-corruption power spectra. The quantity $\Gamma_k = \frac{1}{3} \sum_c \Gamma_{k,c}$ is a motivating abstraction, not a measured physical survivability score.

Under the motivating local low-pass approximation, lower-frequency components would be expected to retain higher detectability on average. We use this expected ordering as an inductive bias rather than as a measured property of the present setup or a guaranteed ordering of the learned branches.

SBR assigns a separate nonlinear estimator to each nominal branch:

$$f_k = \mathcal{E}_{\theta_k}(x_k), \qquad k \in \{L, M, H\}. \tag{11}$$

These estimators can be viewed as amortizing a band-conditioned variational inverse:

$$\widetilde{y}_k = \arg\min_z \frac{1}{2} \sum_{c\in\{R,G,B\}} \int \frac{|M_k(H_c\widehat{z}_c - \widehat{x}_c)|^2}{S_{N,c} + S_{B,c} + \varepsilon_{\text{psd}}}\, d\omega + \mu_k \mathcal{J}_k(z). \tag{12}$$

Here, $H_c(\omega)$ is an unmeasured channel-specific composite transfer function, and this local inverse is a motivating unmeasured abstraction rather than a calibrated physical inversion. The trainable operators' post-optimization frequency responses are not audited in this study.

The width-96 realization uses six bottleneck blocks for LOW, six for MIDDLE, and four for HIGH; the LOW path also has the broadest effective receptive field. This is an architectural allocation across the three nominal branches, not a measured allocation proportional to physical recoverability.

The branch priors $\mathcal{J}_k$ encode different architectural correlation scales: the nominal LOW path aggregates broad spatial support, the MIDDLE path processes the difference branch, and the HIGH path processes the residual complement. This organization is motivated by, but does not calibrate, the physical inverse.

SBR organizes the wall observation into three exactly complementary nominal branches that are processed separately before cross-branch fusion. This construction provides a structured representation of the observed residual cues while preserving the exact complementary sum, without asserting that the branches correspond to independently measured physical frequency bands.

3.3. Content-Conditioned Cross-Band Fusion

CCF integrates the SBR branch features before final image reconstruction. Let the branch set be $\mathcal{B} = \{L, M, H\}$, and let $f_k(\rho)$ denote the feature map at feature-grid coordinate ρ for branch $k \in \mathcal{B}$. The Global Average Pooling descriptors are

$$d_k = \frac{1}{|\Omega_f|} \int_{\Omega_f} f_k(\rho)\, d\rho, \qquad d = d_L \oplus d_M \oplus d_H. \tag{13}$$

The gating network produces observation-conditioned logits

$$\beta = \mathcal{G}_\phi(d), \qquad \beta = (\beta_L, \beta_M, \beta_H). \tag{14}$$

Define

$$\Delta_{\mathcal{B}} = \left\{ \alpha \in \mathbb{R}^3 : \alpha_k \geq 0,\ \sum_{k\in\mathcal{B}} \alpha_k = 1 \right\}. \tag{15}$$

The normalized coefficients are characterized by the entropy-regularized allocation problem

$$\alpha^\star = \arg\max_{\alpha\in\Delta_{\mathcal{B}}} \left[\sum_{k\in\mathcal{B}} \alpha_k \beta_k - \tau \sum_{k\in\mathcal{B}} \alpha_k \log \alpha_k \right], \tag{16}$$

whose solution is

$$\alpha_k^\star = \frac{\exp(\beta_k/\tau)}{\sum_{j\in\mathcal{B}} \exp(\beta_j/\tau)}. \tag{17}$$

We henceforth write $\alpha_k := \alpha_k^\star$ in the synthesis equation. The standard implementation uses $\tau = 1$; Softmax can nevertheless be highly concentrated when the logits differ, and the entropy term does not guarantee balanced weights.

The weighted evidence is synthesized as

$$\begin{aligned} t &= (\alpha_L f_L) \oplus (\alpha_M f_M) \oplus (\alpha_H f_H), \\ \widehat{y} &= x + \mathcal{S}_\psi(t). \end{aligned} \tag{18}$$

The coefficients are observation-conditioned computational coefficients, not optical SNR, calibrated reliability, confidence, or semantic probabilities. CCF receives no Browser, Figure, Game, Software, page-source, or other external category label. The input addition is a normalized-coordinate computational identity skip, not a radiometric identity between wall and screen domains.

In the main width-96 realization, x and the output share a 512×512 grid. The synthesis operator projects the concatenated 288 channels to 192 with a 1×1 convolution and applies two Nonlinear Activation Free (NAF) blocks before RGB prediction. The skip supplies only an identity path for optimization in normalized coordinates; it is not a physical residual base for screen radiance.

All parameters $\Theta = \{\eta_1, \eta_2, \theta_L, \theta_M, \theta_H, \phi, \psi\}$ are learned from paired observations through

$$\begin{aligned} \Theta^\star &= \arg\min_{\Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(\mathcal{R}_\Theta(x), y)], \\ \widehat{y} &= \mathcal{R}_{\Theta^\star}(x). \end{aligned} \tag{19}$$

At inference, x is the only external input; the hidden screen, content label, page source, text, exact geometry, wall reflectance, calibration, and active illumination remain unavailable.

CCF conditions the computational combination of the three branch features on statistics extracted from the current wall observation.

4. Experiments

4.1. Dataset Acquisition and Experimental Protocol

Data were acquired in a controlled indoor no-direct-view setting, as illustrated in Figure 3. A self-luminous display faces a white diffuse relay wall, while a standard RGB camera is positioned behind the display and directed toward the same wall, such that the camera cannot directly observe the content shown on the screen. The room is approximately 5.0 m × 3.6 m; the display-to-wall distance is approximately 1.4 m, and the camera-to-wall distance is approximately 3.5 m. Light emitted by different regions of the screen reaches the wall with different intensities and colors, but spatial propagation and diffuse reflection strongly mix these signals. Consequently, the captured wall image generally appears as a smooth and blurred color footprint in which coarse brightness, color, and layout cues may remain visible, whereas fine text, thin edges, icons, and local textures are much more weakly represented.

Each target screen image was displayed individually, and one RGB image of the corresponding wall footprint was captured as its input. The captured wall image and the original screen image therefore form one input–ground-truth pair. All wall observations are real physical measurements rather than synthetically degraded versions of the target images, and each pair was acquired separately rather than extracted from a video or image burst. The resulting dataset contains 11,700 paired samples divided into four content subsets: Browser, Figure, Game, and Software. Here, Software denotes a dataset content category consisting of screen images from various types of software interfaces, such as plotting, statistical analysis, and data conversion tools, rather than a specific software package; therefore, no single software version applies to this category.

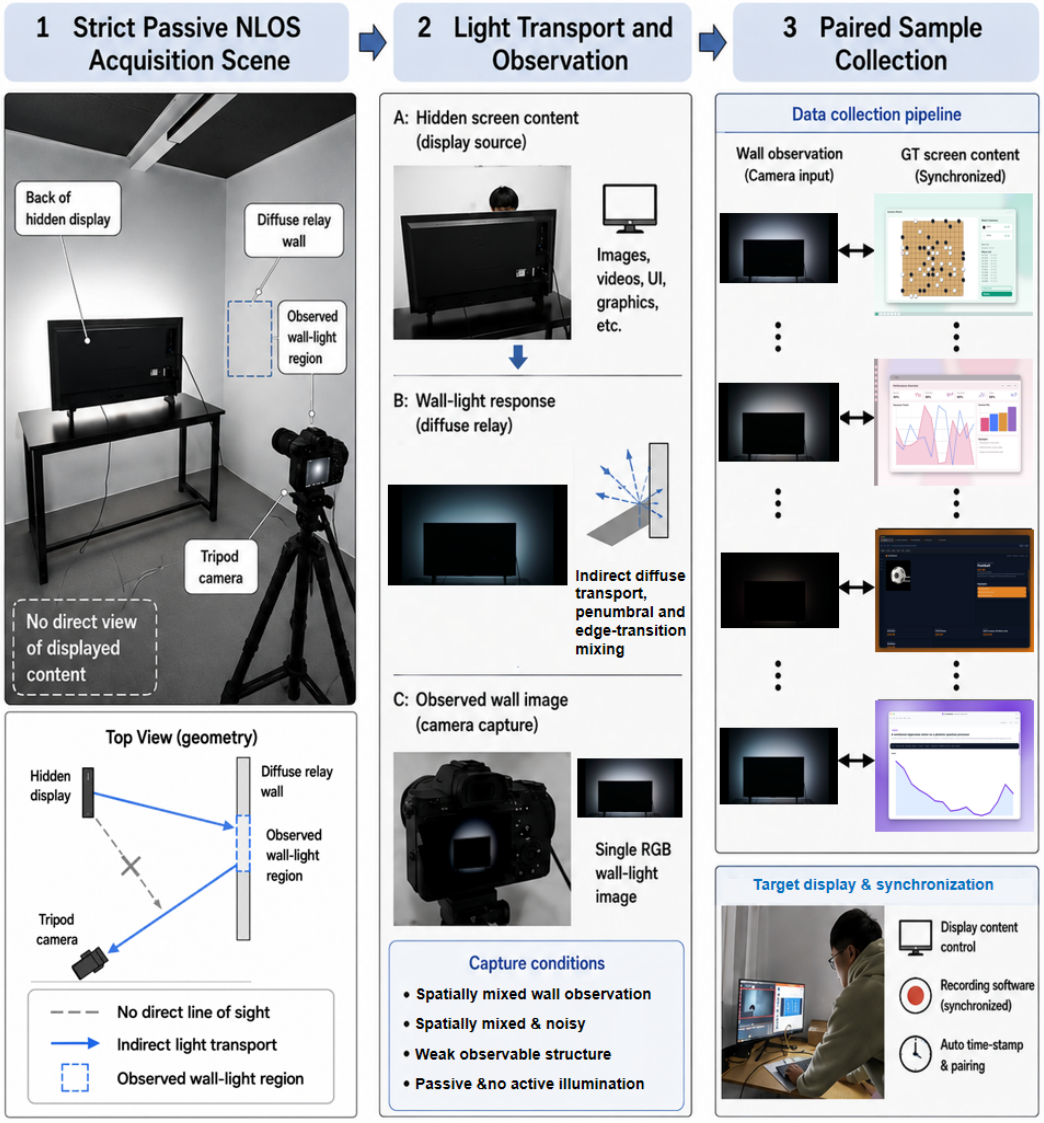


Figure 3. Acquisition and paired-sample construction workflow for the strict passive NLOS dataset. (**Panel 1**) shows the controlled no-direct-view geometry and relay-wall observation region. (**Panel 2**) shows the hidden display, indirect diffuse transport, penumbral and edge-transition mixing, and resulting single-RGB wall observation. These labels are qualitative and do not represent a measured spatial-frequency response or a measured diffraction contribution. (**Panel 3**) shows the operator-side target display, synchronization, and pairing of separately captured wall observations with their target screen images. Arrows and dashed markings indicate the light-transport, viewing, and pairing relationships shown in the figure, while the colors are used for visual distinction and the vertical dots denote omitted intermediate paired samples.

The Browser subset contains approximately 4700 paired samples of webpages and browser interfaces, including navigation bars, text, images, buttons, input fields, icons, and diverse page layouts. This subset mainly evaluates the recovery of coarse webpage organization, major content regions, and the relative arrangement of text and visual elements from blurred wall observations.

The Figure subset contains approximately 1250 paired samples of data visualizations, including line, bar, pie, scatter, and area charts, together with axes, legends, grid lines, and data annotations. Because many important chart elements are thin and locally structured, this subset emphasizes the recovery of overall chart layouts, major curve trends, graphical regions, and color relationships.

The Game subset contains approximately 1250 paired samples of video game content, including interfaces such as Gomoku. Its wall observations are generally darker than those of the other subsets, with an average input brightness of approximately 25–40 compared with approximately 140–220 for the corresponding screen images, providing a challenging setting for reconstruction from weak and low-contrast visual evidence.

The Software subset contains approximately 4500 paired samples of desktop and web applications, editors, dashboards, tables, panels, controls, icons, and dense textual regions. These observations contain more complex imaging variations, such as moiré patterns, mild perspective distortion, glare, nonuniform brightness, and color shifts, and therefore provide a more challenging test of recovering complex interface layouts and local User Interfac (UI) structures.

Figure 4 presents representative input–target pairs from all four subsets. Representative dataset samples are also available at https://github.com/zhengzhiwen123/screen-dataset-samples (accessed on 10 September 2026).

Raw 1920×1080 wall observations and paired targets are independently bicubic-resized to 512×512, converted to RGB, and scaled to [0, 1]. No Region of Interest (ROI), center, or random crop and no mean/std normalization are used. Training augmentation is limited to a paired horizontal flip with probability 0.5; validation and test inputs are not augmented.

The 11,700 paired samples are partitioned deterministically at the pair level into 9955 training, 539 validation, and 1206 test samples. The test set is selected first using a fixed hash-based rule applied to the sample identifier, and the validation set is then selected from the remaining non-test samples using a separate fixed hash-based rule. The random seed is used for training and reproducibility only and does not determine the dataset partition. The current split is pair-level rather than source- or template-grouped.

Training uses AdamW with initial learning rate 2×10^{-4}, weight decay 10^{-4}, batch size 2, and at most 100 epochs. A step-level CosineAnnealingLR schedule is configured with a minimum rate of 10^{-6}; gradients are clipped to 1.0, Automatic Mixed Precision (AMP) is disabled. Early stopping monitors validation reconstruction loss with patience 6 and minimum improvement 10^{-4}. The selected checkpoint is epoch 80 and training stops at epoch 86.

The training objective is

$$\mathcal{L} = \mathcal{L}_{\text{char}} + 0.05\mathcal{L}_{\text{ssim}} + 0.1\mathcal{L}_{\text{band}}, \qquad \mathcal{L}_{\text{band}} = 0.5\|e_L\|_1 + 1.5\|e_M\|_1 + \|e_H\|_1.$$

Here, $\mathcal{L}_{\text{char}}$ averages $\sqrt{e^2 + \epsilon_{\text{char}}^2}$ with $\epsilon_{\text{char}} = 10^{-3}$, and $\mathcal{L}_{\text{ssim}} = 1 - \text{mean}$.

The loss-only branch errors use fixed 7×7, $\sigma = 1.5$, and 15×15, $\sigma = 4.0$, Gaussian decompositions of prediction and target. These fixed loss filters are distinct from the SBR trainable depthwise operators, which use the same kernels only for initialization.

Test metrics are per-image arithmetic means over all 1206 samples. They comprise Peak Signal-to-Noise Ratio (PSNR); Structural Similarity Index Measure (SSIM) [32]; a three-scale mean SSIM obtained by averaging SSIM after zero, one, and two $2\times$ average pooling operations; Learned Perceptual Image Patch Similarity (LPIPS) with AlexNet after mapping $[0, 1]$ to $[-1, 1]$ [33]; and a gradient-magnitude similarity proxy. The latter converts RGB to mean grayscale, uses horizontal and vertical forward differences, and averages $(2g_a g_b + 0.01)/(g_a^2 + g_b^2 + 0.01)$ over pixels.

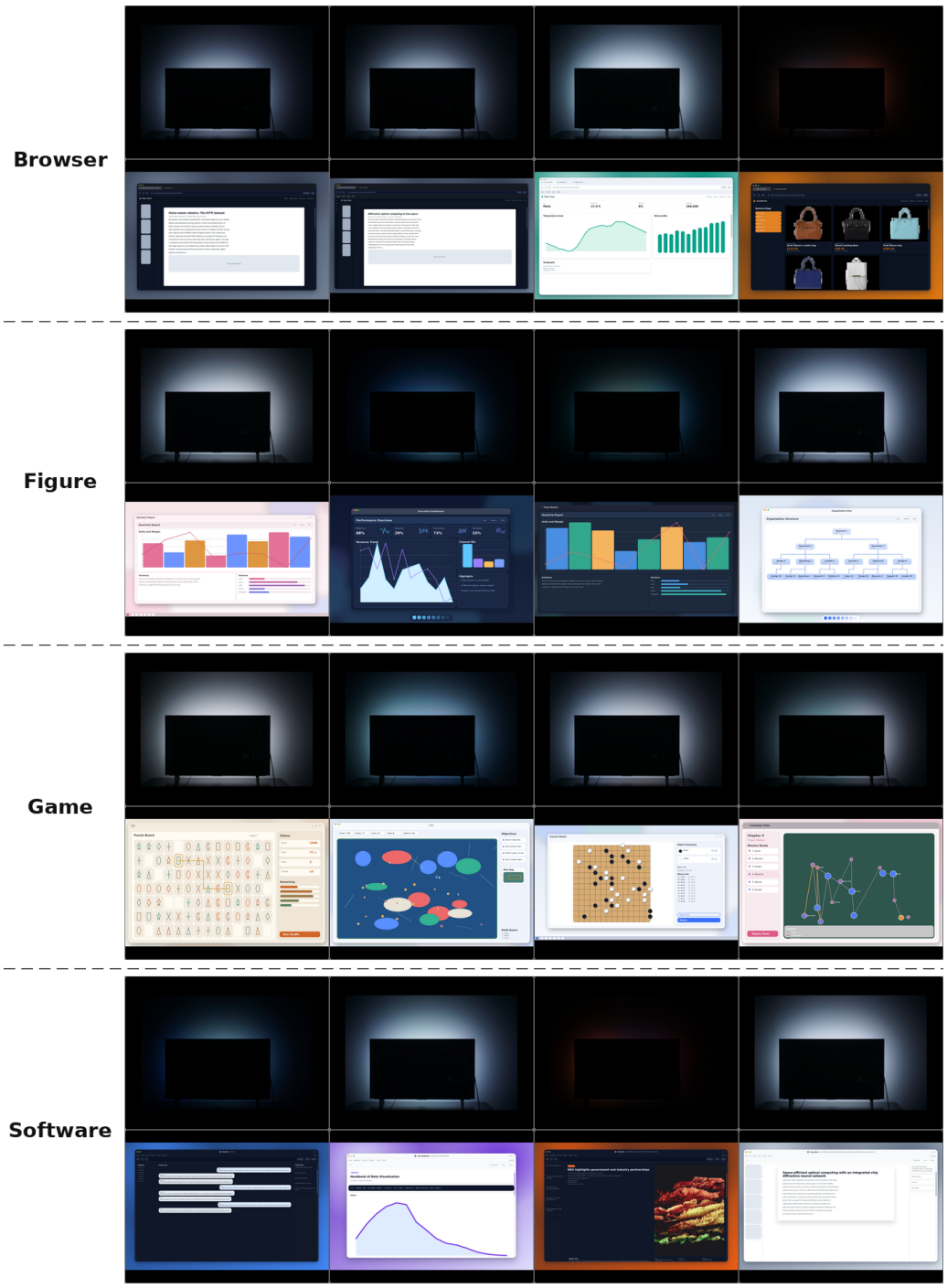


Figure 4. Representative input–target pairs from the Browser, Figure, Game, and Software subsets. Within each subset, the upper row shows physically captured passive NLOS inputs and the lower row shows the synchronized ground-truth screen content. The examples illustrate content-dependent variations in brightness, color, contrast, and structural complexity across the four subsets. The colors reflect the original captured observations and target screen content and do not encode additional categorical or quantitative information.

Our custom three-scale mean averages ordinary SSIM after zero, one, and two 2× average-pooling operations, rather than using the canonical weighted Multi-Scale Structural Similarity Index Measure (MS-SSIM) [34]. The gradient-magnitude quantity is a proxy and does not use phase congruency or the standard Feature Similarity Index (FSIM) weighting [35].

The comparison uses representative task-adapted reconstruction controls under the same paired wall-to-screen task. The recorded comparison is described as using the fixed SGRL partition, a common input resolution, and common evaluation metrics, but the delivery does not include per-baseline manifests sufficient to audit every training recipe and capacity difference. Venue and source labels identify the original publications associated with the compared methods. The controls thus provide an architecture-diverse context rather than a parameter- or compute-matched causal test.

4.2. Overall Quantitative Comparison

Table 1 reports 20.03 dB PSNR, 0.743 SSIM, 0.713 three-scale mean SSIM, 0.410 LPIPS, and 0.946 gradient-magnitude similarity for SGRL on the 1206-image pair-level held-out test set. Among the evaluated controls, SGRL obtains the numerically best value for all five reported metrics, and its PSNR exceeds the recorded AdaRevD value of 12.16 dB by 7.87 dB. These results provide descriptive reconstruction evidence under the present wall-to-screen protocol. Because the evaluated methods are not parameter-, compute-, or training-recipe-matched, the comparison does not isolate the effect of architecture alone or establish relative computational efficiency. Under the evaluated setting, the results nevertheless show that the tested SGRL configuration recovers more of the broad visual structure and local image transitions present in the paired screen targets than the compared controls.

Figure 5 shows the training and validation losses of the SGRL configuration. The validation-loss early-stopping criterion selects the checkpoint at epoch 80, while training terminates at epoch 86. Figure 6 further compares model size and reconstruction quality for SGRL and representative task-adapted reconstruction controls. The scatter adds descriptive model-size context; it does not match parameter count, compute, or training recipes.

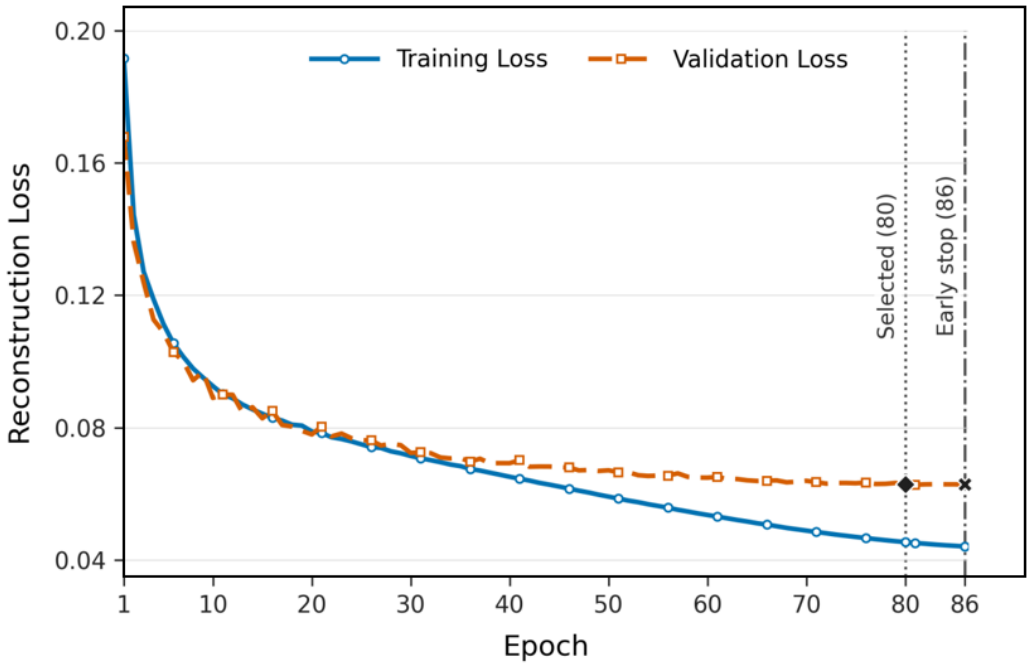


Figure 5. Training convergence of the width-96 SGRL configuration. Training and validation reconstruction losses are shown across epochs. The validation-loss early-stopping criterion selected the checkpoint at epoch 80, and training terminated at epoch 86.

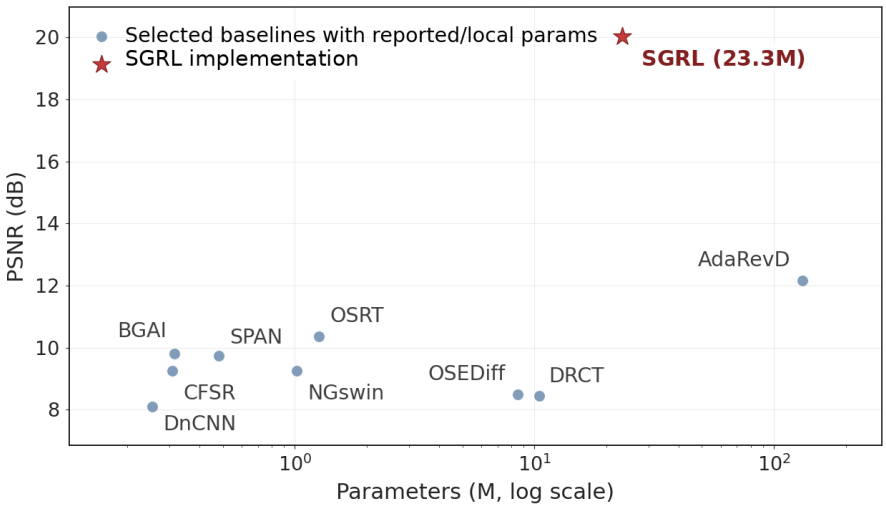


Figure 6. Comparison of parameter count and PSNR between SGRL and the task-adapted reconstruction controls. The horizontal axis shows trainable parameters on a logarithmic scale, and the vertical axis shows PSNR. SGRL is marked by the red star. This comparison provides descriptive model-size context rather than a parameter- or compute-matched efficiency evaluation.

Table 1. Overall comparison with representative task-adapted reconstruction controls, grouped by model family under the common wall-to-screen setting. Higher PSNR, SSIM, 3-scale SSIM, and gradient-magnitude similarity and lower LPIPS indicate better reconstruction. Best and second-best values are highlighted in red and blue, respectively; the ranking is numerical only and does not imply matched training recipes or causal mechanism evidence.

Family	Method	Venue	Overall				
			PSNR ↑	SSIM ↑	3-Scale SSIM ↑	LPIPS ↓	Grad.-Mag. Sim. ↑
Attention/ Transformer	DRCT	CVPRW/ NTIRE'24	8.4500	0.5600	0.4810	0.7240	0.6640
	NGswin	CVPR'23	9.2500	0.5360	0.4260	0.7240	0.6330
	OSRT	CVPR'23	10.3500	0.5590	0.4430	0.7060	0.6190
	SFHformer	ECCV'24	4.4600	0.0200	0.2220	0.9550	0.5200
	BGAI	TIP'26	9.8100	0.5500	0.4480	**0.6980**	0.6340
CNN/ Hybrid	AdaRevD	CVPR'24	**12.1600**	**0.5780**	**0.4860**	0.7170	0.6330
	DnCNN	TIP'17	8.0900	0.4550	0.3850	0.7550	0.5970
	KernelDAN	IJCV'23	8.5000	0.4050	0.3750	0.7880	0.6310
	RealDAN	IJCV'23	8.3900	0.3960	0.3760	0.7910	0.6310
	SPAN	CVPRW/ NTIRE'24	9.7400	0.5080	0.4130	0.7100	0.5990
	CFSR	TIP'24	9.2400	0.4970	0.3960	0.7200	0.5870
Diffusion	DiffBIR	ECCV'24	10.0400	0.5590	0.4450	0.6990	0.6240
	OSEDiff	NeurIPS'24	8.4800	0.5580	**0.4860**	0.7220	**0.6680**
	ResDiff	AAAI'24	9.9700	0.5290	0.4080	0.7350	0.6040
	SinSR	CVPR'24	9.2100	0.4940	0.4450	0.7480	0.6430
Model-based/ Unrolled	DeepSN-Net	TPAMI'25	11.6900	0.5510	0.4630	0.7520	0.6510
Zero-shot	Pixel2Pixel	TPAMI'25	9.4600	0.5050	0.3950	0.7170	0.5920
Ours	SGRL	–	**20.0300**	**0.7430**	**0.7130**	**0.4100**	**0.9460**

Note: The red and blue background colors indicate the best and second-best numerical results, respectively. ↑ indicates that higher values are better, whereas ↓ indicates that lower values are better.

4.3. CCF Weight Analysis

Table 2 summarizes the CCF fusion coefficients on the full 1206-image held-out test set. The learned allocation is strongly concentrated on the HIGH branch, whose mean coefficient is 0.996847 overall and remains between 0.996547 and 0.997243 across the four content categories. The corresponding overall means for MIDDLE and LOW are 0.002739 and 0.000414, respectively. HIGH receives the largest coefficient for all 1206 test observations. The category-wise differences are therefore modest relative to the dominant common

pattern, although the coefficients still exhibit small observation- and category-dependent variation. These values should be interpreted as learned computational fusion coefficients rather than calibrated optical reliability, physical survivability, SNR, or semantic class probabilities. Because CCF acts on branch-specific learned feature maps rather than on raw Fourier-band energies, the dominant HIGH coefficient should not be interpreted as evidence that high-frequency wall transport is physically stronger than low-frequency transport.

Table 2. Mean CCF fusion coefficients on the 1206-image held-out test set, grouped by screen-content category. The coefficients are computational fusion weights predicted from the wall observation and are not calibrated optical reliability or SNR estimates.

Category	Samples	LOW	MIDDLE	HIGH
Browser	535	0.000454	0.003000	0.996547
Figure	131	0.000393	0.002540	0.997066
Game	123	0.000339	0.002418	0.997243
Software	417	0.000391	0.002562	0.997047
Overall	1206	0.000414	0.002739	0.996847

4.4. Per-Content Analysis

Table 3 shows that SGRL achieves the highest PSNR on all four content subsets and overall. It reaches 23.23 dB on Software, 19.84 dB on Figure, 18.62 dB on Game, and 17.92 dB on Browser. Relative to the strongest external control in each subset, the recorded numerical differences are 11.13 dB for Software, 2.33 dB for Figure, 5.87 dB for Game, and 6.55 dB for Browser. These differences are system-level comparisons, not isolated mechanism effects. The score variation also indicates content-dependent recovery, motivating the same-input visual comparison in Figure 7.



Figure 7. Qualitative comparison on eight shared-input examples, with two from each of the Browser, Figure, Game, and Software subsets. Columns show Input, SGRL, AdaRevD, DeepSN-Net, OSRT, DiffBIR, ResDiff, BGAI, and GT. The colored subset labels are used to visually distinguish the four content subsets, while the colors within the reconstruction panels reflect the original image content and do not represent additional quantitative encoding.

Table 3. Per-content PSNR (dB) on the Browser, Software, Figure, and Game subsets. Higher values indicate better reconstruction. Best and second-best results are highlighted in red and blue, respectively; Overall denotes the complete pair-level held-out test set.

Method	Source	Browser	Software	Figure	Game	Overall
DRCT	CVPRW/NTIRE'24	8.4000	7.2800	11.6700	9.2400	8.4500
NGswin	CVPR'23	8.9200	8.4900	11.6400	10.7600	9.2500
OSRT	CVPR'23	9.2200	10.4300	12.5700	12.6200	10.3500
SFHformer	ECCV'24	4.4400	4.9800	3.4100	3.9000	4.4600
BGAI	TIP'26	10.4200	7.7800	11.1100	12.6600	9.8100
AdaRevD	CVPR'24	**11.3700**	11.5200	**17.5100**	12.1100	**12.1600**
DnCNN	TIP'17	7.8800	7.3900	10.6200	8.6600	8.0900
KernelDAN	IJCV'23	8.5000	7.4000	9.6100	11.0300	8.5000
RealDAN	IJCV'23	8.8600	6.6800	10.5400	9.8300	8.3900
SPAN	CVPRW/NTIRE'24	9.5200	9.0900	11.8400	10.6900	9.7400
CFSR	TIP'24	9.1400	8.5900	10.4200	10.6600	9.2400
DiffBIR	ECCV'24	9.2900	9.6600	11.8100	**12.7500**	10.0400
OSEDiff	NeurIPS'24	8.0000	7.4600	13.3000	8.9300	8.4800
ResDiff	AAAI'24	9.3000	9.9600	12.3200	10.4600	9.9700
SinSR	CVPR'24	10.0000	10.0200	3.4100	9.1900	9.2100
DeepSN-Net	TPAMI'25	11.0800	**12.1000**	12.4100	12.1800	11.6900
Pixel2Pixel	TPAMI'25	8.9000	9.2300	11.8600	10.1000	9.4600
SGRL	–	**17.9200**	**23.2300**	**19.8400**	**18.6200**	**20.0300**

Note: The red and blue background colors indicate the best and second-best numerical results, respectively.

SGRL preserves the global layout, major panel boundaries, broad color organization, and coarse chart or diagram structure more consistently. The Game examples remain more strongly affected by weak contrast and low wall-signal strength. Fine text, exact symbols, and small interface elements are still unstable, so the qualitative evidence supports this paper's bounded structure-level claim rather than exact screen-reading.

4.5. Ablation Study

Tables 4 and 5 extend the original deletion-only analysis with removals and substitutions at both the module and mechanism levels. The authors confirm that these experiments used the same dataset, overall experimental protocol, and epoch count. Earlier delivery text stated that the Full row used a different test partition and that some CCF rows came from different fixed checkpoints. Those statements conflict with the authors' common-protocol confirmation and could not be independently reconciled from the supplied records. The present delivery contains no split manifests, per-variant optimization records, initialization records, checkpoint-selection rules, or evaluation-version records with which to verify those details. Protocol provenance is therefore pending, and the tables show single recorded values. Their colors identify only the numerical best and second-best supplied displayed values in each column; they are not an inferential, strictly fair, causal, or statistical ranking.

The recorded values nevertheless expose important local behavior that should not be hidden. For example, FixedGaussian has Figure SSIM 0.7169, whereas SGRL (Full) has 0.7150. IndependentSigmoid has Game LPIPS 0.6059, and SpatialSoftmax has 0.6062; PlainResConv has Figure SSIM 0.7102, and CrossBandAttention has 0.7075. These are descriptive observations only. They do not establish statistical significance, robustness across training seeds, calibrated frequency behavior, or that every difference is caused solely by the named component. If the missing provenance is verified, the same tables can support a common-protocol account of the observed module and internal-mechanism contributions.

Table 4. Overall module-level ablation of SBR and CCF across the Browser, Figure, Game, and Software subsets. All variants follow the same dataset, experimental protocol, and epoch count. Values are reported to four decimal places; a lower LPIPS and higher SSIM and PSNR indicate better reconstruction. Red and blue denote the numerically best and second-best values in each metric column, respectively.

Variant	Browser			Figure			Game			Software		
	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑
w/o SBR	0.6839	0.4894	12.2998	0.6135	0.6627	17.9056	0.7077	0.4957	13.9457	0.6002	0.5947	13.4983
SBR → WaveletRep	0.6477	0.5246	13.6634	0.6216	0.6517	18.0186	0.7085	0.5123	14.9129	0.5521	0.6335	15.3381
SBR → FourierRep	0.6496	0.5288	13.4911	0.6259	0.6479	17.8869	0.7132	0.5057	14.6755	0.5633	0.6285	14.9916
w/o CCF	0.5608	0.6210	15.6167	0.6175	0.6866	18.4568	0.6652	0.5963	16.6797	0.4226	0.7560	19.0474
CCF → SEConcat	0.5156	0.6680	16.8809	0.6081	0.7126	19.3780	**0.6256**	**0.6404**	**17.8402**	0.3695	0.8065	21.0997
CCF → CBAMConcat	**0.5118**	**0.6719**	**17.0147**	**0.6063**	**0.7136**	**19.4545**	0.6270	0.6400	17.7823	**0.3669**	**0.8114**	**21.2120**
SGRL (Full)	**0.4386**	**0.6957**	**17.9167**	**0.5429**	**0.7150**	**19.8369**	**0.5723**	**0.6481**	**18.6246**	**0.2832**	**0.8395**	**23.2270**

Note: The red and blue background colors indicate the best and second-best numerical results, respectively. ↑ indicates that higher values are better, whereas ↓ indicates that lower values are better.

Table 5. Internal-mechanism ablation of SBR and CCF across the Browser, Figure, Game, and Software subsets under the same experimental protocol. Values are reported to four decimal places; a lower LPIPS and higher SSIM and PSNR indicate better reconstruction. Red and blue denote the numerically best and second-best values in each metric column, respectively. "Adaptive weight prediction" refers to the mechanism previously labeled "reliability estimation".

Variant	Browser			Figure			Game			Software		
	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑
w/o complementary decomposition	0.4982	0.6718	17.1483	0.5951	0.7122	19.5770	0.6119	0.6433	**18.0416**	0.3580	0.8099	21.2645
decomposition → FixedGaussian	**0.4896**	**0.6806**	**17.2520**	0.6024	0.6967	18.7561	0.6120	0.6437	17.9060	**0.3492**	**0.8171**	**21.5036**
decomposition → UnconstrainedFilterbank	0.4980	0.6717	17.1858	0.5959	0.7111	19.5751	0.6125	0.6402	18.0110	0.3545	0.8122	21.4601
w/o band-specific estimators	0.6735	0.4894	13.1653	0.6231	0.6519	18.0942	0.7088	0.5021	14.7755	0.5748	0.6065	14.7919
band-specific estimators → SharedEstimator	0.5122	0.6576	16.7860	0.5988	0.7072	19.3081	0.6090	0.6358	17.8491	0.3659	0.8011	20.8905
band-specific estimators → HomogeneousEstimators	0.4997	0.6661	17.0482	0.5940	0.7113	19.4439	0.6096	0.6399	17.9788	0.3600	0.8073	21.2646

Table 5. *Cont.*

Variant	Browser			Figure			Game			Software		
	LPIPS ↓	**SSIM ↑**	**PSNR ↑**	**LPIPS ↓**	**SSIM ↑**	**PSNR ↑**	**LPIPS ↓**	**SSIM ↑**	**PSNR ↑**	**LPIPS ↓**	**SSIM ↑**	**PSNR ↑**
w/o adaptive weight prediction	0.4990	0.6695	17.1180	0.5953	0.7115	19.4545	0.6136	0.6403	17.9615	0.3611	0.8077	21.1683
adaptive weight prediction → IndependentSigmoid	0.4992	0.6704	17.1045	0.5941	0.7115	19.5603	**0.6059**	0.6405	17.9945	0.3587	0.8088	21.3623
adaptive weight prediction → SpatialSoftmax	0.4951	0.6746	17.2201	**0.5907**	**0.7141**	**19.5984**	0.6062	**0.6447**	18.0376	0.3544	0.8134	21.4911
w/o cross-band synthesis	0.5441	0.6317	16.1092	0.6149	0.6849	18.5035	0.6448	0.6123	17.2398	0.4106	0.7599	19.4089
cross-band synthesis → PlainResConv	0.5299	0.6607	16.7897	0.6135	0.7102	19.3272	0.6349	0.6343	17.8107	0.3815	0.8008	20.9452
cross-band synthesis → CrossBandAttention	0.5064	0.6643	16.9952	0.5975	0.7075	19.4525	0.6067	0.6364	17.9190	0.3660	0.8028	21.1434
SGRL (Full)	**0.4386**	**0.6957**	**17.9167**	**0.5429**	**0.7150**	**19.8369**	**0.5723**	**0.6481**	**18.6246**	**0.2832**	**0.8395**	**23.2270**

Note: The red and blue background colors indicate the best and second-best numerical results, respectively. ↑ indicates that higher values are better, whereas ↓ indicates that lower values are better.

At the module level, removing SBR or CCF, or replacing them with the tested conventional representation and fusion alternatives, consistently reduces reconstruction performance across the four content subsets. This pattern supports the complementary roles of the two modules in the hidden-screen reconstruction task: SBR provides a structured way to represent wall-side visual evidence that has been degraded to different degrees, whereas CCF allows these complementary cues to be combined according to the current observation rather than with a fixed fusion strategy. At the internal-mechanism level, the full SGRL configuration is numerically best in 11 of the 12 reported subset-level metric comparisons, with Figure SSIM as the only exception. Taken together, the recorded ablations are consistent with the intended design principle that differently degraded wall-side cues should first be represented separately and then adaptively combined for reconstructing the hidden screen, while the provenance limitations stated above preclude stronger causal or statistical claims.

The experiments provide three distinct levels of network evidence. The full-system comparison reports integrated SGRL performance; the two recorded ablation tables describe module- and mechanism-level behavior subject to the pending provenance above; and the following common-U-Net study separately examines the proposed bottleneck representation. These levels are complementary but not interchangeable, and none is a physical calibration of spatial-frequency survivability.

4.6. Bottleneck Representation Comparison

To complement the component ablations, we evaluate the proposed representation within a common U-Net encoder–decoder framework. The bottleneck is instantiated with ten candidate representations: plain convolution, a residual block, a NAF-style block, Fourier and wavelet representations, a SwinIR-style residual Swin Transformer block [12], a Uformer-style window representation [13], Restormer-style multi-Dconv head transposed attention [11], a MambaIR-like representation, and Expanded-SGRLRep-Ours. This auxiliary comparison focuses on the observed behavior of these bottleneck configurations. The authors confirm the same dataset, overall protocol, and epoch count for all candidates, but the exact split identifiers, per-candidate configurations, and checkpoint-selection records are not present in this delivery. Thus, the comparison has the same reported epoch budget but is not parameter- or compute-matched, and its protocol provenance remains pending. It should not be interpreted as an efficiency benchmark or a capacity-controlled causal attribution.

Table 6 reports all three metrics for the four content categories. Within the supplied recorded values, Expanded-SGRLRep-Ours is numerically lower in LPIPS and higher in SSIM and PSNR than the other nine candidates across all 12 category-specific metric entries. Its LPIPS, SSIM, and PSNR results are 0.5838, 0.6061, and 15.6619 for Browser; 0.6288, 0.6898, and 18.7272 for Figure; 0.6789, 0.5991, and 17.0161 for Game; and 0.4426, 0.7510, and 19.0045 for Software. The cell colors visualize only this numerical ordering among the supplied displayed values. They do not remove the pending protocol provenance or establish strict fairness, causality, statistical significance, or parameter/compute matching.

Figure 8 summarizes arithmetic differences from the same recorded values by selecting the numerically strongest remaining candidate separately at each position, rather than one fixed baseline. Across Browser, Figure, Game, and Software, the PSNR gains are 1.1843, 0.5573, 1.0838, and 1.8512 dB; the SSIM gains are 0.0368, 0.0170, 0.0473, and 0.0456; and the LPIPS reductions are 0.0443, 0.0013, 0.0365, and 0.0513, respectively. All displayed differences are positive, including the small Figure LPIPS difference. Because the underlying provenance remains pending, these differences are descriptive rather than inferential.

Table 6. Comparison of ten bottleneck representations within a common U-Net encoder–decoder framework on the Browser, Figure, Game, and Software subsets. A higher PSNR and SSIM and lower LPIPS indicate better reconstruction. All values are reported to four decimal places. Red and blue cells identify only the numerical best and second-best supplied displayed values in each metric column; protocol provenance remains pending, so the colors do not establish a strictly fair, causal, or statistical comparison. The candidates have the same author-reported epoch budget but are not parameter- or compute-matched, and their values are not results of the complete SGRL model.

Representation	Browser			Figure			Game			Software		
	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑
Reduced-PlainConvRep	0.6389	0.5629	**14.4776**	0.6368	0.6615	18.0667	0.7183	0.5470	**15.9323**	0.5115	0.6904	16.9571
Reduced-ResBlockRep	0.6462	0.5644	14.2898	0.6365	0.6688	**18.1699**	0.7311	0.5434	15.6954	0.5096	0.6910	16.8495
Reduced-NAFBlockRep	0.6434	0.5578	14.2330	0.6330	0.6633	18.0571	0.7219	0.5371	15.5865	0.5199	0.6804	16.5321
Reduced-FourierRep	0.6472	0.5578	14.3370	**0.6301**	0.6663	18.1476	0.7245	0.5361	15.7494	0.5180	0.6911	16.8424
Reduced-WaveletRep	0.6471	0.5516	14.1363	0.6347	0.6600	17.9224	0.7308	0.5234	15.4563	0.5262	0.6753	16.4446
Reduced-SwinIRRSTBRep	**0.6281**	**0.5693**	14.4775	0.6355	0.6657	18.0336	0.7218	0.5410	15.6234	**0.4939**	0.7051	**17.1533**
Reduced-UformerWindowRep	0.6287	0.5690	14.3621	0.6332	**0.6728**	18.1674	**0.7154**	**0.5518**	15.8240	0.4976	**0.7054**	17.1403
Reduced-RestormerMDTARep	0.6514	0.5482	13.8528	0.6351	0.6603	17.9466	0.7437	0.5154	15.1008	0.5281	0.6719	16.2803
Reduced-MambaIRLikeRep	0.6565	0.5478	14.0757	0.6351	0.6636	18.0744	0.7350	0.5313	15.2976	0.5232	0.6821	16.5757
Expanded-SGRLRep-Ours	**0.5838**	**0.6061**	**15.6619**	**0.6288**	**0.6898**	**18.7272**	**0.6789**	**0.5991**	**17.0161**	**0.4426**	**0.7510**	**19.0045**

Note: The red and blue background colors indicate the best and second-best numerical results, respectively. ↑ indicates that higher values are better, whereas ↓ indicates that lower values are better.

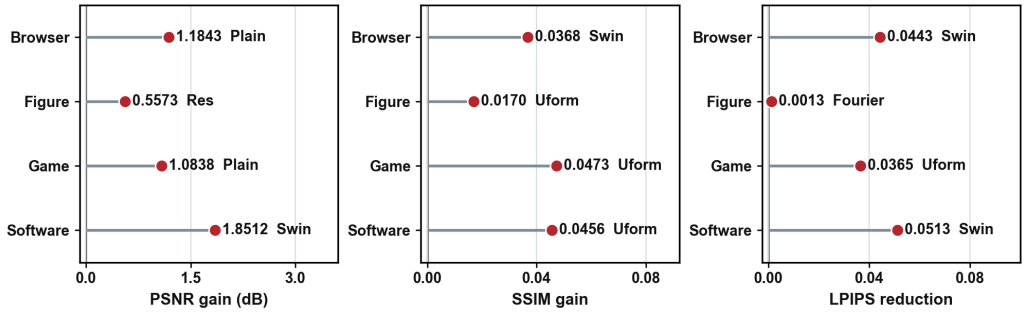


Figure 8. Descriptive per-category margin of Expanded-SGRLRep-Ours relative to the numerically strongest recorded alternative in Table 6. For each category and metric, the proposed representation is excluded before selecting the strongest of the other nine candidates. PSNR gain and SSIM gain are computed as the proposed result minus the highest competing result, whereas LPIPS reduction is computed as the lowest competing LPIPS minus the proposed result; all margins are shown in the original metric units. These are arithmetic differences from the supplied single recorded values; protocol provenance is pending and the plot is not a parameter- or compute-matched inference. Plain, Res, Swin, Uform, and Fourier denote Reduced-PlainConvRep, Reduced-ResBlockRep, Reduced-SwinIRRSTBRep, Reduced-UformerWindowRep, and Reduced-FourierRep, respectively.

Subject to verification of the common-protocol records, these results would support the bounded conclusion that the proposed representation performs best within this tested ten-candidate U-Net setting. They must not replace the main SGRL results: the complete SGRL model combines SBR, CCF, and the RGB reconstruction path, whereas this auxiliary experiment evaluates bottleneck representations in a common U-Net scaffold. The same epoch count does not make the candidates parameter- or compute-matched, and the results do not establish efficiency or the superiority of the complete SGRL system over these U-Net variants. Having characterized the reconstruction components and bottleneck representations, we next evaluate finite-pool candidate narrowing using the complete SGRL outputs.

4.7. Candidate-Identification Controls

The candidate-identification analyses use all 1206 paired samples in the pair-level held-out test set. The same held-out samples used for the main reconstruction-quality evaluation are reused for the downstream candidate-identification analyses while remaining strictly disjoint from the data used to train SGRL. These experiments evaluate whether the information retained in the SGRL reconstructions can support identification of the paired screen target within explicitly defined finite candidate pools. They therefore characterize bounded finite-pool ranking rather than open-set screen-content recognition.

For query q_i, let y_i denote its paired ground-truth target, and let $\mathcal{C}_i$ denote a candidate set of size K containing y_i as the unique positive candidate and $K-1$ negative candidates. A fixed similarity function $\mathrm{sim}(q_i, y_{i,j})$ ranks every candidate item $y_{i,j} \in \mathcal{C}_i$, and r_i denotes the rank of the paired target after sorting candidates from most to least similar. For a test set of N queries, Top-k, Mean Reciprocal Rank (MRR), and median rank are defined as

$$\text{Top-}k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[r_i \le k]. \tag{20}$$

$$\mathrm{MRR} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{r_i}, \qquad \mathrm{MedianRank} = \mathrm{median}_i(r_i). \tag{21}$$

These metrics characterize ranking within explicitly defined finite candidate pools rather than open-set screen-content recognition.

The principal finite-pool setting is a 16-way candidate-identification task using all 1206 held-out SGRL outputs. Here, "16-way" means that each query is ranked against 16 candidate ground-truth screen images: one is its paired target, and the remaining 15 are negative targets drawn without replacement from the other 1205 held-out test samples. Thus, every query forms a ranking problem containing one correct target and 15 distractors.

One deterministic representation used for candidate ranking is the 4116-D Privacy feature descriptor with cosine similarity. The descriptor jointly represents coarse brightness, spatial color, and global chromatic statistics. It concatenates a 32×32 grayscale intensity map (1024-D), a 32×32 RGB map (3072-D), and a 20-D LAB chromatic histogram formed by 10-bin a^* and b^* histograms, giving $1024 + 3072 + 20 = 4116$ dimensions. This descriptor is used only for downstream candidate ranking and is not part of the SGRL network or its training procedure.

To reduce dependence on any single downstream representation, the automatic-jury experiment evaluates four candidate judges together with one score-level ensemble, all using the same seed 16-way candidate pool for each SGRL query. Layoutis a deterministic 1600-D handcrafted judge emphasizing downsampled intensity and coarse spatial-layout information. Privacy feature uses the deterministic 4116-D descriptor defined above to compare coarse brightness, spatial color, and global chromatic information. Color-edge is a deterministic 2320-D handcrafted judge emphasizing color-distribution and edge-structure information. Contrastive is a learned 256-D two-tower judge that maps SGRL reconstructions and candidate ground-truth targets into a shared embedding space for similarity ranking. Ensemble introduces no additional image descriptor and directly sums the raw scores produced by Layout, Privacy feature, Color-edge, and Contrastive without learned fusion weights.

The Layout, Privacy feature, and Color-edge judges contain no learned or pretrained parameters. The learned contrastive judge is trained on 2400 training pairs (600 per content family) for 900 steps with batch size 256, temperature 0.07, and bidirectional cross-entropy. This training is performed only for the downstream evaluator and does not train or modify SGRL. The judge-training data are strictly disjoint from the 1206 held-out candidate-identification queries. In the documented run, no validation split or early-stopping stage is used, and a single checkpoint is saved after the fixed 900-step schedule before post-training embedding evaluation. Reconstructions use the wall tower and candidate targets use the Ground-Truth (GT) tower. Non-image metadata, including filenames, file paths, content labels, and candidate indices, does not enter the similarity score.

Table 7 evaluates whether the information retained in the SGRL reconstructions can support identification of the paired screen target within the bounded candidate pool. Under the 16-way setting, completely random ranking has an expected Top-1 rate of only $1/16 = 6.25\%$. In comparison, the Privacy feature judge reaches 38.80%, 83.60%, and 96.10% Top-1/5/10, respectively, while the Layout judge reaches 38.90% Top-1. The Ensemble reaches 36.33%, 82.17%, and 96.00% Top-1/5/10. These results indicate that the information retained in SGRL reconstructions is not limited to overall pixel or perceptual similarity; it also contains discriminative cues that can help distinguish the paired screen target from alternative candidates within the defined finite pool.

The different judges should not be interpreted as competing reconstruction methods. They provide different downstream views of the same fixed SGRL outputs, and their recorded rates therefore reflect the dependence of candidate identification on the chosen feature representation and scoring rule. In particular, the deterministic handcrafted judges record higher rates than the learned contrastive judge under the documented protocol. The purpose of this experiment is therefore not to establish superiority among the judges but to examine whether reconstructed information remains useful for bounded target identifica-

tion under multiple downstream representations. In this sense, candidate identification provides a functional complement to PSNR, SSIM, and LPIPS, which primarily measure reconstruction similarity at the pixel, structural, and perceptual levels.

Table 7. Automatic blind-jury candidate-identification results for SGRL under the shared 16-way random-all protocol. Each of 1206 queries contains one paired target and 15 sampled negatives, with the same candidate set used by all judges. Top-k rates are reported as percentages.

Automatic Jury	Top-1	Top-5	Top-10
Ensemble	36.33%	82.17%	96.00%
Layout	38.90%	81.70%	91.80%
Privacy feature	38.80%	83.60%	96.10%
Color-edge	25.30%	61.70%	86.00%
Contrastive	7.80%	33.60%	62.70%

We additionally evaluate how candidate identification changes as the candidate pool becomes larger using a distinct shared-pool protocol.

For each query i, pool size K, and candidate seed ν, one candidate-ID list $\mathcal{C}_{i,K,\nu}$ is generated and reused unchanged for wall input, U-Net-lite, Plain Convolutional Neural Network (CNN), NAF-lite, and SGRL output. All five query sources therefore use aligned queries, identical positive and negative candidate IDs, the same 1328-D descriptor, and cosine similarity; only the query source changes. The 1328-D descriptor combines 16×16 RGB, a 48-D LAB histogram, layout, and edge maps. Random-all pools are generated independently across K over seeds but are shared across the five query sources within each (i, K, ν). For hard-nearest evaluation, one deterministic candidate-side list per (i, K) is likewise reused by all five query sources.

Figure 9 shows that the recorded Top-5 rate generally decreases as the candidate budget increases. Under random-all sampling, the SGRL curve is numerically highest from $K = 8$ to $K = 128$. Under hard-nearest sampling, all query sources degrade more sharply, and their relative differences are smaller at several candidate budgets. These results are descriptive paired-pool observations, and no statistical-significance claim is assigned.

The 4116-D automatic-jury setting and the 1328-D candidate-budget setting are distinct downstream ranking protocols and should not be treated as interchangeable measurements. Neither descriptor changes the SGRL network itself. Absolute candidate-identification rates remain dependent on the downstream representation, candidate-pool composition, random seed, and candidate budget. Accordingly, the evidence in this section supports bounded finite-pool candidate identification and candidate narrowing rather than open-set screen-content recognition or a real-world attack success rate. The stability of this observation under changes in candidate sampling is examined separately in Section 4.8.

4.8. Candidate-Identification Evaluation Settings and Stability Checks

Having established in Section 4.7 that SGRL reconstructions retain information that supports bounded candidate identification, we next examine whether this observation depends strongly on a particular random draw of negative candidates. Table 8 summarizes several candidate-identification settings designed to separate changes in candidate sampling, the downstream descriptor, candidate budget, gallery size, and evaluation protocol. The rows therefore serve as protocol and stability checks rather than as directly ranked competing methods.

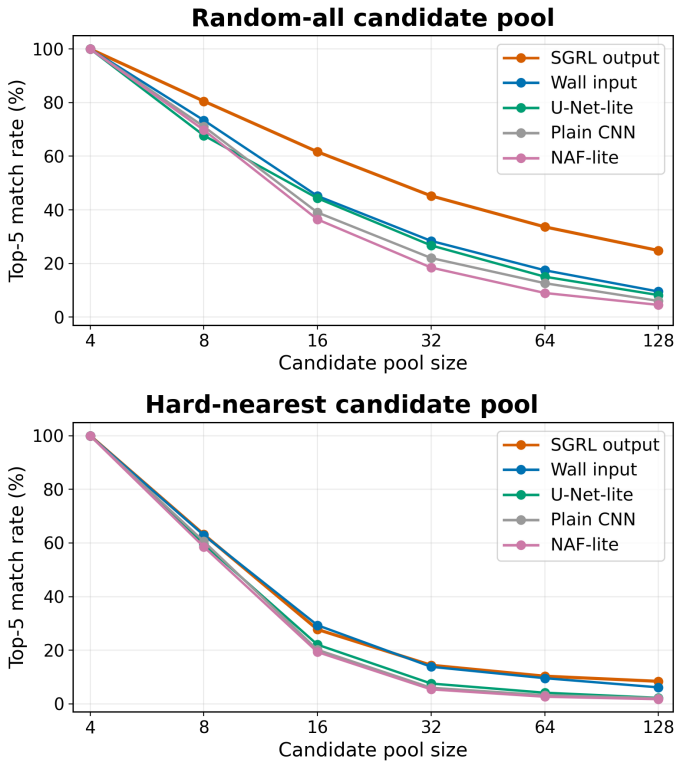


Figure 9. Top-5 candidate-identification performance as candidate-pool size increases over $K \in \{4, 8, 16, 32, 64, 128\}$. For every query, K, and seed, wall input, U-Net-lite, Plain CNN, NAF-lite, and SGRL output use exactly the same candidate IDs, 1328-D descriptor, and cosine-similarity rule. Random-all reports the mean over 20 paired candidate-pool resamplings. Hard-nearest uses a shared deterministic candidate list built from same-family candidates; Game at $K = 128$ requires five cross-family negatives because only 123 Game samples are available. The curves measure paired finite-pool matching, not open-set recognition or repeated-training uncertainty.

Table 8. Candidate-identification settings and stability checks. Rows use different downstream descriptors, sampling schemes, and candidate budgets, and are therefore not ranked across protocols. Top-k rates are reported as percentages.

Evidence Block	Protocol Summary	Top-1	Top-5	Top-10	MRR	Median Rank
16-way single-seed *a*	SGRL output; 4116-D Privacy feature; random-all; cosine; seed *a*	38.58%	85.25%	96.00%	0.5733	2
16-way single-seed *b*	SGRL output; 4116-D Privacy feature; random-all; cosine; seed *b*	40.25%	84.17%	96.00%	0.5768	2
Automatic blind jury	SGRL output; random-all; 16-way; four-judge score fusion	36.33%	82.17%	96.00%	0.5495	2
Repeated candidate resampling	SGRL output; 1328-D Privacy feature; random-all; 16-way; 30 candidate-pool resamplings	27.29%	61.25%	80.58%	0.4334	4
Defense-sensitivity reference	SGRL reconstruction; 720 held-out test GTs; 720-way full gallery; 2816-D handcrafted descriptor; cosine similarity	4.30%	16.40%	24.40%	–	–

The first two rows of Table 8 provide a direct candidate-sampling check under the same 16-way protocol. Both use the same 1206 cached SGRL reconstructions, the same 4116-D Privacy feature defined in Section 4.7, the same cosine-similarity ranking rule, and the same candidate budget of one paired target plus 15 negative targets. The only intended difference is the random seed used to select the negative candidates: seed *a* and seed *b*. Their Top-1 rates are 38.58% and 40.25%, respectively, while both reach 96.00%

Top-10. Their median ranks are both 2. Thus, changing this single candidate-pool draw does not cause the bounded identification effect to disappear. This two-seed comparison is descriptive, however, and is not treated as a statistical robustness test by itself.

The automatic-blind-jury row reports the score-level ensemble introduced in Section 4.7. It uses the same 1206 SGRL queries and 16-way finite-pool interpretation but a different downstream scoring construction, combining the raw scores of the Layout, Privacy feature, Color-edge, and Contrastive judges without learned fusion weights. Its 36.33% Top-1, 82.17% Top-5, and 96.00% Top-10 rates provide another view of the same fixed SGRL reconstructions. Because its downstream scoring construction differs from the single 4116-D Privacy-feature protocol, these numbers are not used to rank one protocol against the other.

For a more systematic sampling stability check, the repeated-resampling experiment fixes all 1206 queries, the trained SGRL reconstructions, the 1328-D descriptor, and the 16-way candidate budget, while changing only the negative candidate pools. For every query, one paired target is retained and 15 negative candidates are independently resampled from the other 1205 held-out targets across 30 random seeds. The resulting mean Top-1, Top-5, and Top-10 rates are 27.29%, 61.25%, and 80.58%, respectively, with an MRR of 0.4334 and a median rank of 4.

Let σ_{seed} denote the seed-level standard deviation of a metric over N_ν independent candidate seeds. The two-sided normal-approximation 95% confidence intervals for the seed-level means are computed as

$$\bar{x} \pm 1.96 \frac{\sigma_{\text{seed}}}{\sqrt{N_\nu}}. \tag{22}$$

giving $[26.90, 27.67]$%, $[60.98, 61.53]$%, and $[80.38, 80.78]$% for Top-1, Top-5, and Top-10, respectively. These relatively concentrated intervals quantify the variability caused by negative-candidate resampling while the trained SGRL, query set, descriptor, and candidate budget remain fixed. They therefore show that the observed finite-pool identification behavior is not attributable only to one particularly favorable draw of negative candidates. These intervals do not quantify uncertainty from repeated SGRL training.

The related shared-pool candidate-budget experiment in Figure 9 uses the same 1328-D descriptor construction and 20 paired candidate-pool resamplings. At $K = 16$, SGRL records 61.65% Top-5 in that experiment, compared with 61.25% in the separate SGRL-only 30-resampling evaluation. These close values provide descriptive evidence that the same finite-pool candidate-ranking signal remains visible under two independently instantiated resampling protocols; they are not used as a formal equivalence test.

Table 8 additionally includes a 720-way full-gallery defense-sensitivity reference. Whereas a 16-way query ranks one paired target against 15 negative candidates, the 720-way setting ranks each reconstruction against a fixed gallery of 720 held-out ground-truth screen images containing one paired positive and 719 negatives. It therefore represents a substantially larger candidate space. Under its documented 2816-D handcrafted descriptor and cosine-similarity rule, the Top-1, Top-5, and Top-10 rates are 4.30%, 16.40%, and 24.40%, respectively. For reference, a random Top-1 ranking in a 720-way gallery has an expected rate of only $1/720 \approx 0.14\%$.

The 720-way result must nevertheless be interpreted within its own protocol. It uses a different gallery size, downstream descriptor, and evaluation setting from the 16-way rows. Consequently, the difference between 4.30% and the 38.58–40.25% 16-way Top-1 values does not represent a directly comparable performance drop, and the rows in Table 8 should not be numerically ranked across protocols.

Taken together with Table 7, these stability checks provide a functional complement to the reconstruction-quality evaluation. Table 7 shows that fixed SGRL reconstructions retain information that can support identification of the paired screen target within a bounded finite candidate pool. Table 8 further shows that this observation remains visible when candidate-sampling seeds are changed and when negative candidate pools are repeatedly resampled. At the same time, the absolute Top-k rates depend on the downstream descriptor, candidate pool, random seed, candidate budget, and gallery size. Accordingly, these experiments support bounded finite-pool candidate identification and candidate narrowing rather than open-set screen-content recognition or a real-world attack success rate.

4.9. Digital Perturbation Sensitivity for Candidate Identification

This analysis evaluates finite-pool candidate-identification sensitivity under post-acquisition digital perturbations using a separately trained width-56 auxiliary reconstruction model operating at 128 × 128 resolution. A fixed set of 720 wall observations and their corresponding paired targets is used for queries and a 720-way candidate gallery. Each recorded wall observation is digitally transformed before reconstruction, after which a fixed 2816-D downstream descriptor and cosine similarity are used for candidate ranking. Every perturbation condition is evaluated deterministically on the same 720 sample IDs.

This candidate-identification analysis is distinct from the reconstruction-quality evaluation in Section 4.10, which uses the primary width-96, epoch-80 checkpoint at 512 × 512 resolution and reports direct PSNR and SSIM. The two analyses share the same 720 sample IDs and the five selected transformation definitions used in the reconstruction-quality study, but they differ in their model, checkpoint, resolution, reconstruction output, metric implementation, and evaluation endpoint. Figure 10 summarizes the candidate-identification response under the digital controls. These post-acquisition transformations characterize digital sensitivity and do not constitute physical re-acquisition or validated defense mechanisms.

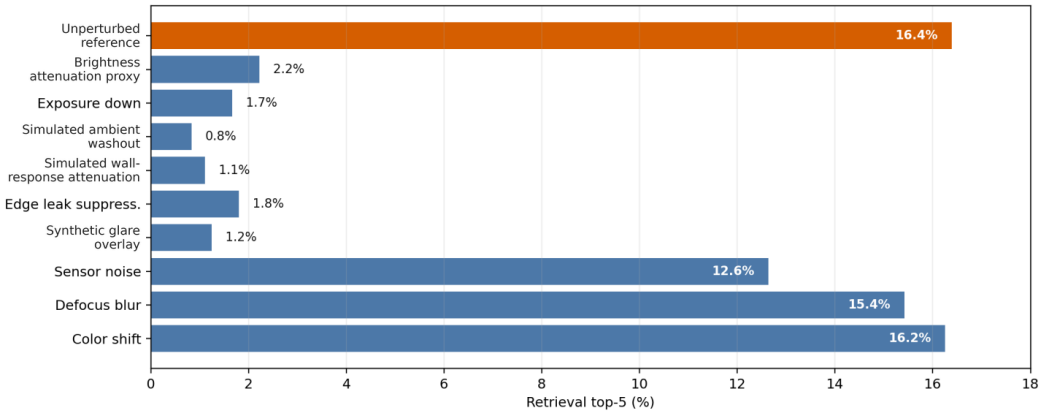


Figure 10. Candidate-identification response under post-acquisition digital perturbation controls, including the unperturbed reference, brightness attenuation, exposure reduction, simulated ambient washout, simulated wall-response attenuation, edge-response attenuation, synthetic glare overlay, sensor noise, defocus blur, and color scaling. All transformations are applied digitally to previously recorded wall observations before reconstruction and candidate ranking. The results characterize the digital sensitivity of the finite-pool identification pipeline and do not represent physical re-acquisition or validated defense mechanisms.

The fixed subset comprises 247 Browser, 131 Figure, 123 Game, and 219 Software pairs. Each digitally modified recorded wall input passes through SGRL before retrieval against the same 720 paired targets. The fixed 2816-D downstream descriptor concatenates 32 × 32 low-frequency grayscale, 16 × 16 RGB, and 32 × 32 edge features, followed by mean centering, L2 normalization, and cosine ranking. All conditions share the query

subset, gallery, descriptor, and ranking rule; only the perturbation condition changes, and non-image metadata is excluded. The 2816-D descriptor is used only for downstream ranking and is not part of the SGRL network.

The deterministic input transformations are the following: brightness $x \mapsto 0.55x$; exposure $x \mapsto 0.45(x - 0.02)$; washout $x \mapsto 0.45x + 0.28$; RGB wall attenuation $[0.32, 0.34, 0.36]$; edge-response attenuation using the 60th-percentile edge threshold, Gaussian $\sigma = 6$, and attenuation 0.65; defocus with $\sigma = 3.0$; sensor noise with $\sigma = 0.055$ and seed; glare centered at $(0.60W, 0.35H)$, scale $0.15W$, and amplitude 0.45; and color scaling $[1.18, 0.88, 0.74]$. Each reported condition is one run over the same $n = 720$ samples.

Ambient washout gives the largest reduction in Table 9, changing the Top-1/5/10 from 4.3%/16.4%/24.4% to 0.1%/0.8%/1.4%; wall-response attenuation also reduces retrieval strongly. These within-setting changes show digital sensitivity, not an effective physical defense. That claim would require physical re-acquisition under modified illumination, display, relay surface, or camera conditions.

Table 9. Post-acquisition digital sensitivity controls for finite-pool candidate identification on the fixed 720-query subset and corresponding 720-way gallery. Within this auxiliary candidate-identification experiment, all deterministic conditions use the same reconstruction model, 2816-D downstream descriptor, and cosine ranking rule, while only the digital transformation applied to the recorded wall observation is changed. Lower Top-*k* values indicate reduced candidate-identification performance relative to the unperturbed reference. These results characterize digital sensitivity and are not physical re-acquisition measurements or validated defense evaluations. Colors are used only to distinguish the different perturbation conditions and have no additional quantitative meaning.

Condition	Top-1 ↓	Top-5 ↓	Top-10 ↓	Interpretation
Unperturbed reference	4.3000%	16.4000%	24.4000%	720-way reference
Simulated ambient washout	0.1000%	0.8000%	1.4000%	Largest reduction in candidate identification
Brightness attenuation	0.4000%	2.2000%	3.9000%	Reduced-input-intensity condition
Synthetic glare overlay	0.3000%	1.2000%	3.1000%	Simulated photometric interference
Simulated wall-response attenuation	0.1000%	1.1000%	2.2000%	Proxy for reduced relay response

Note: ↓ indicates that lower values are better.

The digital measurements characterize post-acquisition sensitivity of the SGRL reconstruction under a controlled downstream evaluation; they do not extend the physical reconstruction claims.

4.10. Reconstruction Quality Under Digital Perturbations

We additionally evaluated reconstruction quality under post-acquisition digital perturbations using the primary width-96, epoch-80 checkpoint on a fixed set of 720 sample IDs. Table 10 reports the clean baseline and the corresponding PSNR and SSIM changes. Relative to the clean mean of 19.947414 dB and 0.736079, ambient washout and wall-response attenuation produce the largest mean reductions, namely −13.723264 dB and −0.417893, and −11.613967 dB and −0.389513, respectively. Defocus blur produces the smallest mean reduction, −1.111215 dB and −0.019252, under the tested parameterization. PSNR and SSIM do not necessarily rank perturbation sensitivity identically. These measurements are post-acquisition digital sensitivity tests rather than a validation of physical defense mechanisms.

The original Top-*k* analysis uses a separately trained width-56 auxiliary model at 128 × 128 resolution. The present reconstruction-quality analysis uses the primary width-96, epoch-80 checkpoint at 512 × 512 resolution and reports direct PSNR/SSIM. The two analyses share the exact 720 sample IDs and the five selected transformation definitions

only; they do not share their model, checkpoint, resolution, reconstruction outputs, metric implementation, or endpoint.

Table 10. Post-acquisition digital sensitivity analysis on 720 fixed sample IDs using the primary width-96, epoch-80 checkpoint. Values are sample means. Δ values are relative to the clean reconstruction mean. The perturbations are digital transformations of recorded wall observations and do not represent physical re-capture defenses.

Condition	*N*	PSNR (dB)	SSIM	ΔPSNR/ΔSSIM
Clean	720	19.947414	0.736079	–
Brightness attenuation	720	13.685523	0.545527	−6.261891/−0.190552
Ambient washout	720	6.224151	0.318186	−13.723264/−0.417893
Wall−response attenuation	720	8.333447	0.346566	−11.613967/−0.389513
Sensor noise	720	17.647804	0.490066	−2.299610/−0.246013
Defocus blur	720	18.836199	0.716827	−1.111215/−0.019252

4.11. Wall-Observation Intensity Stratification

We further examined reconstruction quality across rank-based wall-observation mean-intensity strata. For each canonically preprocessed RGB wall image $x_i \in [0,1]^{3\times H_{\text{img}}\times W_{\text{img}}}$, we defined

$$B_i = \frac{1}{3H_{\text{img}}W_{\text{img}}} \sum_{c,h,w} x_i(c,h,w).$$

The 1206 test samples were sorted by B_i and divided into Low, Medium, and High tertiles of 402 samples each. This quantity is the normalized digital mean of the recorded wall observation; it is not calibrated luminance and does not denote the display's nominal brightness.

Table 11 shows a non-monotonic association between these strata and reconstruction quality. The Low group has mean intensity 0.03104101 and achieves 20.681651 dB PSNR and 0.749141 SSIM; the Medium group has mean intensity 0.12315934 and achieves 17.739247 dB and 0.683052; and the High group has mean intensity 0.17368067 and achieves 21.680098 dB and 0.795734. Thus, the observed order is Medium < Low < High for both metrics, rather than a monotonic increase with intensity.

The category composition also differs across the tertiles: Low contains 187 Browser, 35 Figure, 33 Game, and 147 Software samples; Medium contains 210, 2, 66, and 124, respectively; and High contains 138, 94, 24, and 146. We therefore treat the result as a descriptive stratification of this test set. The aggregate differences do not isolate a causal effect of wall-observation intensity, and the mean intensity should not be confused with display luminance.

Table 11. Reconstruction quality across rank-based wall-observation mean-intensity tertiles on the full 1206-sample test set. Each group contains 402 samples. Intensity is the normalized digital mean of the recorded RGB wall observation; it is not display luminance.

Group	*N*	Mean Intensity	PSNR (dB)	SSIM	Intensity Range
Low	402	0.03104101	20.681651	0.749141	0.00433480–0.09506804
Medium	402	0.12315934	17.739247	0.683052	0.09528255–0.15579195
High	402	0.17368067	21.680098	0.795734	0.15595090–0.18894343

5. Discussion

5.1. Interpretation of the Main Findings

The experimental results support the feasibility of structure-level hidden-screen reconstruction from a single passive diffuse-wall observation under the evaluated controlled indoor geometry. Across the overall and content-specific evaluations, the recovered images preserve broad brightness distribution, region layout, and major visual structure more effectively than the evaluated reconstruction controls. This result should be interpreted as a recovery of coarse and intermediate visual organization rather than an exact inversion of the hidden display; fine text, small symbols, and other weakly supported details remain constrained by the information lost during indirect light transport.

The CCF analysis further clarifies the learned behavior of SGRL. The HIGH branch receives the largest fusion coefficient for all 1206 held-out observations, while LOW and MIDDLE receive much smaller coefficients. This concentration does not imply that high spatial frequencies are physically transmitted more reliably through the wall. CCF operates on branch-specific learned feature maps rather than raw Fourier-band energies, and its coefficients are computational fusion variables rather than calibrated measurements of optical SNR, physical survivability, or confidence. The result therefore indicates a stable learned preference within the tested network rather than a measured frequency-domain property of the display–wall–camera path.

The additional sensitivity analyses provide complementary evidence about the conditions under which the reconstruction changes. Post-acquisition digital perturbations reduce reconstruction quality to different degrees, with ambient washout and wall-response attenuation producing the largest reductions among the tested transformations, while defocus blur produces a smaller reduction under the selected parameters. The wall-observation intensity analysis is non-monotonic: the Medium-intensity tertile performs below both the Low- and High-intensity tertiles. Because the content composition also differs among these groups, the observed association should be treated as descriptive and does not establish a causal dependence of reconstruction quality on wall intensity alone.

5.2. Relation to Prior Work and Method Positioning

The present setting differs from conventional active NLOS imaging, where controlled illumination and often time-resolved measurements provide additional transport information for recovering hidden geometry or reflectance [1,2]. Here, reconstruction uses only one steady-state RGB photograph of a diffuse relay wall, without active illumination, transient sensing, or a calibrated transport model. Passive NLOS approaches based on accidental pinholes, visible corners, relay-surface observations, and blank-wall evidence [3–6] operate under different geometric and sensing assumptions. Our contribution is therefore narrower: we examine a paired wall-to-screen inverse under a controlled hidden-display configuration and test whether the indirect wall observation retains sufficient evidence for structure-level screen reconstruction.

The task also differs from conventional image restoration. Methods such as Restormer, SwinIR, and Uformer [11–13] generally reconstruct a clean image from a degraded observation that remains in approximately the same image domain. In the present problem, the observed wall image and the hidden screen target belong to different domains and are connected by an unknown indirect optical transport process. SGRL addresses this cross-domain setting by using SBR to construct complementary representation paths and CCF to condition their computational combination on the current observation. The numerical comparisons in Table 1 provide reconstruction context under the common wall-to-screen task, but they should not be interpreted as a parameter-, compute-, or training-recipe-matched causal comparison.

5.3. Limitations and Future Work

The main physical evidence is limited to one controlled indoor geometry, one diffuse wall observation per target, and a pair-level data split. Website, project, game scene, figure source, and visual template are not used as grouping constraints, so visually related layouts may occur across the training, validation, and test partitions. The reported results therefore characterize within-domain reconstruction under the present split and do not establish source- or template-disjoint generalization. Cross-room, cross-wall, cross-camera, cross-display, distance, exposure, and outdoor generalization also remain untested.

The comparison experiments have additional methodological limitations. The external reconstruction controls and replacement candidates are not parameter- or compute-matched, and the reported ablation and replacement results do not quantify repeated-training variation or statistical significance. Consequently, the numerical rankings describe the recorded experimental configurations rather than isolating the causal contribution of architecture or computational budget. Future evaluations should include repeated training, stronger protocol matching, and source- or template-disjoint partitions.

The physical interpretation of the learned representation also remains bounded. The post-training spatial-frequency responses of the SBR operators were not directly measured against the display–wall–camera transfer path, and no controlled optical frequency-response experiment was performed. The nominal frequency ordering is therefore a design motivation rather than an empirically calibrated wall transfer function. Likewise, CCF coefficients are not calibrated against optical SNR, attenuation, or physical confidence, and the isolated contribution of the direct computational input skip was not separately evaluated. Future work should combine measured optical-transfer characterization with cross-condition physical acquisition to determine how the learned representation changes across different environments and imaging systems.

Finally, the recovered visual structure does not imply exact small-text, password, or character reading. The candidate-identification experiments use finite candidate pools and depend on the downstream descriptor, sampling procedure, seed, and candidate budget; they therefore do not establish open-set recognition or real-world attack success. Similarly, the digital perturbation experiments operate on previously recorded observations and do not substitute for physical re-acquisition or validate a physical defense. Future work should evaluate candidate identification under source-disjoint galleries and test physical mitigation strategies through controlled re-capture.

5.4. Ethical and Dual-Use Considerations

This optical side-channel study has dual-use implications. It is intended to characterize unintended display-information leakage under controlled research conditions rather than to provide a deployment guide for surveillance. Research data and code should exclude credentials, identifiers, and private content, respect applicable source licenses, and be used only with appropriate authorization and consent. The digital sensitivity experiments should likewise be interpreted as diagnostic analyses rather than validated physical defenses; mitigation claims require controlled physical re-acquisition under modified display, illumination, relay-wall, and camera conditions.

6. Conclusions

This study addresses bounded passive hidden-screen reconstruction from one ordinary RGB photograph of a diffuse wall when the screen is outside the camera's direct view.SGRL combines the nominal complementary construction paths of SBR with observation-conditioned CCF, whose weights are computational mixing coefficients rather than calibrated optical reliability measures.

In the controlled indoor protocol, the method achieves 20.03 dB PSNR and 0.743 SSIM on 1206 paired held-out samples, supporting structure-level recovery of broad visual organization while not establishing exact text recovery, open-set recognition, or universal NLOS generalization.

The CCF, digital-sensitivity, and wall-intensity analyses further characterize the tested computational behavior but do not substitute for optical calibration or physical re-acquisition.

Future work should evaluate source- and template-disjoint data, repeated training, cross-room/wall/camera/display acquisition, and measured optical transfer responses to determine how stable the present bounded result is beyond the controlled setting.

Author Contributions: Conceptualization, J.Z. and Z.Z.; methodology, J.Z. and Z.Z.; software, J.Z.; validation, J.Z. and W.T.; formal analysis, J.Z.; investigation, J.Z. and X.H.; resources, Z.Z. and X.H.; data curation, J.Z. and W.T.; writing—original draft preparation, J.Z.; writing—review and editing, Z.Z., W.T., X.H. and H.Z.; visualization, J.Z. and W.T.; supervision, Z.Z. and H.Z.; project administration, Z.Z.; funding acquisition, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Representative samples from the dataset used in this study are publicly available at https://github.com/zhengzhiwen123/screen-dataset-samples(accessed on 10 September 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Velten, A.; Willwacher, T.; Gupta, O.; Veeraraghavan, A.; Bawendi, M.G.; Raskar, R. Recovering Three-Dimensional Shape Around a Corner Using Ultrafast Time-of-Flight Imaging. *Nat. Commun.* **2012**, *3*, 745. [CrossRef]
2. O'Toole, M.; Lindell, D.B.; Wetzstein, G. Confocal Non-Line-of-Sight Imaging Based on the Light-Cone Transform. *Nature* **2018**, *555*, 338–341. [CrossRef]
3. Torralba, A.; Freeman, W.T. Accidental Pinhole and Pinspeck Cameras: Revealing the Scene Outside the Picture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2012.
4. Bouman, K.L.; Ye, V.; Yedidia, A.B.; Durand, F.; Wornell, G.W.; Torralba, A.; Freeman, W.T. Turning Corners into Cameras: Principles and Methods. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2017.
5. Saunders, C.; Murray-Bruce, J.; Goyal, V.K. Computational Periscopy with an Ordinary Digital Camera. *Nature* **2019**, *565*, 472–475. [CrossRef]
6. Sharma, P.; Aittala, M.; Schechner, Y.Y.; Torralba, A.; Wornell, G.W.; Freeman, W.T.; Durand, F. What You Can Learn by Staring at a Blank Wall. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021.
7. Loughry, J.; Umphress, D.A. Information Leakage from Optical Emanations. *ACM Trans. Inf. Syst. Secur.* **2002**, *5*, 262–289. [CrossRef]
8. Kuhn, M.G. Optical Time-Domain Eavesdropping Risks of CRT Displays. In *Proceedings of the IEEE Symposium on Security and Privacy*; IEEE: Piscataway, NJ, USA, 2002; pp. 3–18.
9. Backes, M.; Durmuth, M.; Unruh, D. Compromising Reflections, or How to Read LCD Monitors around the Corner. In *Proceedings of the IEEE Symposium on Security and Privacy*; IEEE: Piscataway, NJ, USA, 2008.
10. Nassi, B.; Pirutin, Y.; Swisa, R.; Shamir, A.; Elovici, Y.; Zadov, B. Lamphone: Passive Sound Recovery from a Desk Lamp's Light Bulb Vibrations. In Proceedings of the USENIX Security Symposium, Boston, MA, USA, 10–12 August 2022.
11. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H.; Shao, L. Restormer: Efficient Transformer for High-Resolution Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2022.
12. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. SwinIR: Image Restoration Using Swin Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*; IEEE: Piscataway, NJ, USA, 2021.
13. Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; Li, H. Uformer: A General U-Shaped Transformer for Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2022.

14. Li, B.; Liu, X.; Hu, P.; Wu, Z.; Lv, J.; Peng, X. All-In-One Image Restoration for Unknown Corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2022.
15. Potlapalli, V.; Zamir, S.W.; Khan, S.; Khan, F.S. PromptIR: Prompting for All-in-One Image Restoration. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023.
16. Hsu, C.C.; Lee, C.M.; Chou, Y.S. DRCT: Saving Image Super-Resolution Away from Information Bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*; IEEE: Piscataway, NJ, USA, 2024; pp. 6133–6142.
17. Choi, H.; Lee, J.; Yang, J. N-Gram in Swin Transformers for Efficient Lightweight Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2023; pp. 2071–2081.
18. Yu, F.; Wang, X.; Cao, M.; Li, G.; Shan, Y.; Dong, C. OSRT: Omnidirectional Image Super-Resolution with Distortion-Aware Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2023; pp. 13283–13292.
19. Jiang, X.; Zhang, X.; Gao, N.; Deng, Y. When Fast Fourier Transform Meets Transformer for Image Restoration. In Proceedings of the Computer Vision–ECCV, Milan, Italy, 29 September–4 October 2024; pp. 381–402. [CrossRef]
20. Wang, Q.; Mao, Y.; Guo, R.; Wang, M.; Wei, J.; Pan, H. Broadcast-Gated Attention With Identity Adaptive Integration for Efficient Image Super-Resolution. *IEEE Trans. Image Process.* **2026**, *35*, 1408–1422. [CrossRef]
21. Mao, X.; Li, Q.; Wang, Y. AdaRevD: Adaptive Patch Exiting Reversible Decoder Pushes the Limit of Image Deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 25681–25690. [CrossRef]
22. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155. [CrossRef]
23. Luo, Z.; Huang, Y.; Li, S.; Wang, L.; Tan, T. End-to-End Alternating Optimization for Real-World Blind Super Resolution. *Int. J. Comput. Vis.* **2023**, *131*, 3152–3169. [CrossRef]
24. Wan, C.; Yu, H.; Li, Z.; Chen, Y.; Zou, Y.; Liu, Y.; Yin, X.; Zuo, K. Swift Parameter-Free Attention Network for Efficient Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*; IEEE: Piscataway, NJ, USA, 2024; pp. 6246–6256.
25. Wu, G.; Jiang, J.; Jiang, J.; Liu, X. Transforming Image Super-Resolution: A ConvFormer-Based Efficient Approach. *IEEE Trans. Image Process.* **2024**, *33*, 6071–6082. [CrossRef]
26. Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Dai, B.; Yu, F.; Ouyang, W.; Qiao, Y.; Dong, C. DiffBIR: Toward Blind Image Restoration with Generative Diffusion Prior. In Proceedings of the Computer Vision–ECCV, Milan, Italy, 29 September–4 October 2024; pp. 430–448. [CrossRef]
27. Wu, R.; Sun, L.; Ma, Z.; Zhang, L. One-Step Effective Diffusion Network for Real-World Image Super-Resolution. In *Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA; Volume 37, pp. 92529–92553.
28. Shang, S.; Shan, Z.; Liu, G.; Wang, L.; Wang, X.; Zhang, Z.; Zhang, J. ResDiff: Combining CNN and Diffusion Model for Image Super-Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA; Volume 38, pp. 8975–8983. [CrossRef]
29. Wang, Y.; Yang, W.; Chen, X.; Wang, Y.; Guo, L.; Chau, L.P.; Liu, Z.; Qiao, Y.; Kot, A.C.; Wen, B. SinSR: Diffusion-Based Image Super-Resolution in a Single Step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 25796–25805.
30. Deng, X.; Zhang, C.; Jiang, L.; Xia, J.; Xu, M. DeepSN-Net: Deep Semi-Smooth Newton Driven Network for Blind Image Restoration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 2632–2646. [CrossRef]
31. Ma, Q.; Jiang, J.; Zhou, X.; Liang, P.; Liu, X.; Ma, J. Pixel2Pixel: A Pixelwise Approach for Zero-Shot Single Image Denoising. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 4614–4629. [CrossRef]
32. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
33. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 586–595.
34. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale Structural Similarity for Image Quality Assessment. In Proceedings of the Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 9–12 November 2003; pp. 1398–1402.
35. Zhang, L.; Zhang, L.; Mou, X.; Zhang, D. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Trans. Image Process.* **2011**, *20*, 2378–2386. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.