MAKING DARKNESS CONTAGIOUS: PRIOR-GUIDED SYNTHESIS FOR LOW-LIGHT OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-light object detection is constrained by scarce annotations, while preset synthesis rules can miss the degradation combinations found in real target images. We propose *Contagious Degradation Synthesis* (**CoDeS**), which learns reusable degradation "strains" from unlabeled low-light images and transfers them to labeled normal-light scenes. CoDeS represents target appearance with seven physically motivated descriptors and trains a conditional variational autoencoder (CVAE) to learn their joint variation conditioned on discovered illumination modes. Sampled configurations guide an image-formation-inspired online renderer that generates varied low-light views while preserving source boxes and categories. We further introduce a lightweight *Low-Light Feature Adapter* (**LFA**) and jointly train it with the SAM 3 detector on these views using inherited source annotations. Our approach requires no target training annotations or paired normal-light/low-light images, and synthesis is confined to training. Experiments on ExDark, NOD, and DARK FACE show significant improvements over pretrained SAM 3. Comparisons with alternative synthesis methods further show that CoDeS generates images that more closely match the local appearance and luminance statistics of real low-light data.

1 INTRODUCTION

When illumination drops, object categories remain unchanged, but the visual cues needed to recognize and localize them become less reliable. Detectors trained on abundant normal-light data consequently struggle with low-light inputs (Hong et al., 2021; 2024). A direct solution is to collect and annotate more low-light images, yet poor visibility makes object boundaries difficult to identify and reliable labeling costly (Du et al., 2024). Covering diverse lighting conditions and camera characteristics further increases this burden. This motivates adapting detectors to low light by reusing the scene diversity and object annotations already available in normal-light datasets.

Low-light synthesis provides a practical route to this goal. Camera-based pipelines simulate exposure changes, sensor noise, and image processing while retaining the annotations of normal-light images (Hong et al., 2021; Cui et al., 2021). However, specifying how an image can degrade does not determine which degradation conditions are representative of a target domain. Real low-light scenes range from uniformly dim environments to deep shadows surrounding isolated bright regions (Sun et al., 2025), with associated variations in noise, blur, and color. In methods such as MAET (Cui et al., 2021) and ISP-Teacher (Zhang et al., 2024), key synthesis parameters are drawn from prescribed distributions. These sampling rules can miss the combinations of degradation factors found in real target images. Effective synthesis therefore requires learning how these factors vary together under different low-light conditions.

Our key insight is that existing low-light images can serve as reusable sources of degradation information, even when their objects are unannotated. Their observable appearance provides a basis for learning characteristic combinations of illumination, noise, blur and color changes. Once represented in a compact parameter space, these combinations can be sampled and transferred across many labeled scenes. Real low-light images thus provide appearance information, while normal-light datasets provide diverse scene content and reliable supervision. This motivates making low-light degradation *contagious*: learned degradation patterns can "infect" normal-light images while

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

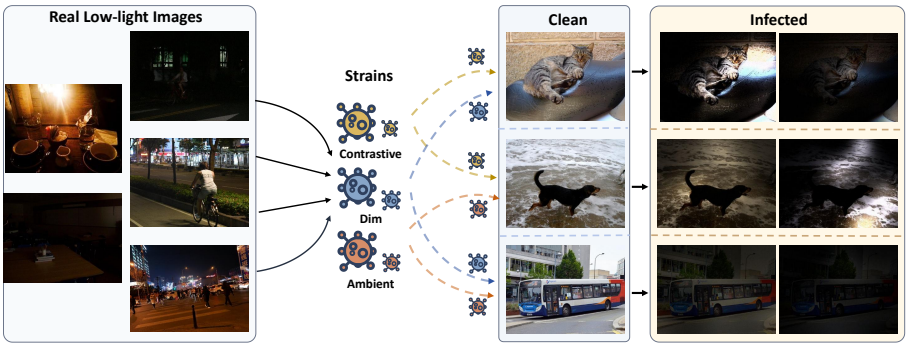


Figure 1: CoDeS learns degradation patterns (Strains) from unlabeled real low-light images and transfers them to labeled normal-light images (Clean). The resulting synthetic low-light images (Infected) exhibit diverse degradations while preserving source object annotations.

preserving their object geometry and annotations. Each labeled scene can then yield multiple low-light training examples without repeated image acquisition or labeling.

Following this idea, we propose *Contagious Degradation Synthesis* (**CoDeS**), illustrated in Figure 1. We describe each unlabeled target image using seven physically motivated appearance measurements: brightness, shadow contrast, visible noise, blur, red–green color bias, blue–green color bias, and spatial illumination variation. These measurements connect observable degradation effects to interpretable synthesis controls. We group images by brightness and shadow contrast to identify illumination modes, which define different degradation "strains". A conditional variational autoencoder (CVAE) then learns the joint variation of all seven measurements within each mode. Samples from this learned prior guide an image-formation-inspired online renderer that applies the degradations to normal-light images while retaining their object boxes and categories. Repeated sampling exposes each labeled scene to varied combinations of low-light degradations throughout training.

To translate this training variation into low-light detection capability, we introduce a lightweight *Low-Light Feature Adapter* (**LFA**) for pretrained SAM 3 (Carion et al., 2026). Inserted between the frozen image encoder and detector, the LFA combines bottleneck channel transformations with depth-wise spatial aggregation to perform detection-oriented feature adaptation. We jointly optimize the LFA and detector using the inherited source annotations, while keeping the image and text encoders frozen. Existing unlabeled target images are used to fit the degradation prior, while no target detection annotations or paired normal-light and low-light images are required. At inference, the adapted detector processes real low-light images directly, with the LFA active and the synthesis pipeline omitted.

Our main contributions are summarized as follows:

- We propose CoDeS, which learns a transferable degradation prior from existing unlabeled low-light images. Seven physically motivated appearance measurements and an illumination-conditioned CVAE capture the joint variation of degradation factors across distinct illumination modes.
- We develop an online synthesis pipeline that transfers sampled degradation configurations to labeled normal-light scenes. It preserves source annotations and generates varied low-light training views without additional low-light image acquisition or annotation.
- We combine the proposed low-light synthesis strategy with the LFA to adapt SAM 3 to low-light conditions. Experiments on multiple benchmarks demonstrate substantial improvements in its ability to recognize and localize objects under low light.

2 Related Work

2.1 Low-Light Image Synthesis

Low-light synthesis expands training data by changing the appearance of labeled normal-light images while retaining their object annotations. One approach uses photometric augmentation to produce varied dark views. For example, Kennerley et al. (2023) introduce NightAug, which combines

random brightness, contrast, gamma, blur, and noise transformations with local image variations. These operations are flexible, but their combinations are not constrained by a camera imaging model.

Camera-based synthesis instead models exposure reduction, sensor noise, and image signal processing (ISP) (Hong et al., 2021; Cui et al., 2021; Zhang et al., 2024). This provides interpretable control over image degradation and can preserve physical relationships, such as the dependence of noise on signal intensity. However, key degradation parameters in MAET and ISP-Teacher are drawn from prescribed distributions. These rules describe how images can degrade, but do not necessarily reflect how often different conditions occur together in a particular target domain. Physical plausibility alone therefore does not ensure that synthetic data covers the variations found in real low-light images.

2.2 Learning for Low-Light Object Detection

One line of research improves the visual cues available to the detector. Image enhancement increases visibility, but its objectives may favor visual quality over the details needed for detection (Hong et al., 2021; Shin et al., 2026). This has encouraged methods that learn features for the detection task. They enhance hierarchical features (Hashmi et al., 2023) or reduce illumination effects through color relationships in the spatial and frequency domains (Hong et al., 2024; Sun et al., 2025). Their effectiveness nevertheless depends in part on how well the training data represents real low-light conditions.

Another line uses auxiliary tasks to strengthen detector training. Image recovery and transformation prediction extract additional supervision from synthetic image pairs (Hong et al., 2021; Cui et al., 2021). SGLDet derives a reconstruction target from the low-light input through enhancement, denoising, and Fourier fusion (Shin et al., 2026). Auxiliary learning can also support transfer without real target images. DAI-Net combines Retinex-based learning with synthetic low-light training in this source-only setting (Du et al., 2024). Without target observations, however, training cannot directly account for differences between simulated and target conditions.

When unlabeled target images are available, domain adaptation can address this mismatch directly. Teacher–student methods use target predictions as pseudo-labels and learn from consistency across models and image views (Kennerley et al., 2023; Zhang et al., 2024). Their reliance on predictions makes pseudo-label errors a concern, motivating calibration and bias correction (Cui et al., 2025). Another strategy aligns source and target domains at the image and feature levels. Hu et al. (2026) propose DLDA, which combines luminance-aware image translation with adversarial feature alignment for this purpose. These strategies mainly use target images to guide appearance transfer or detector learning. A separate challenge is to learn which degradation conditions should be generated when expanding labeled source data. This requires accounting for both the range of each factor and how the factors vary together in real low-light images.

3 Methodology

3.1 Overview

CoDeS learns reusable low-light degradation patterns and transfers them to labeled normal-light scenes for detector adaptation. Given a labeled source dataset $\mathcal{D}_S = \{(x_i^S, y_i^S)\}$ and an unlabeled low-light target dataset $\mathcal{U}_T = \{x_i^T\}$, our goal is to adapt SAM 3 (Carion et al., 2026) to low-light detection using the source object annotations. As shown in Figure 2, we characterize target appearance with seven physically motivated descriptors and train a conditional variational autoencoder (CVAE) to learn their joint distribution conditioned on discovered illumination modes. Its shared latent representation captures how degradation factors vary together within different low-light "strains". Sampling this prior provides coordinated configurations for an online renderer, which generates diverse low-light views of source scenes while preserving their boxes and categories. We jointly train the Low-Light Feature Adapter (LFA) and detector on these views, keeping the image and text encoders frozen. Target detection labels and paired normal-light/low-light images are unnecessary. And target validation and test images are excluded from prior fitting and detector training.

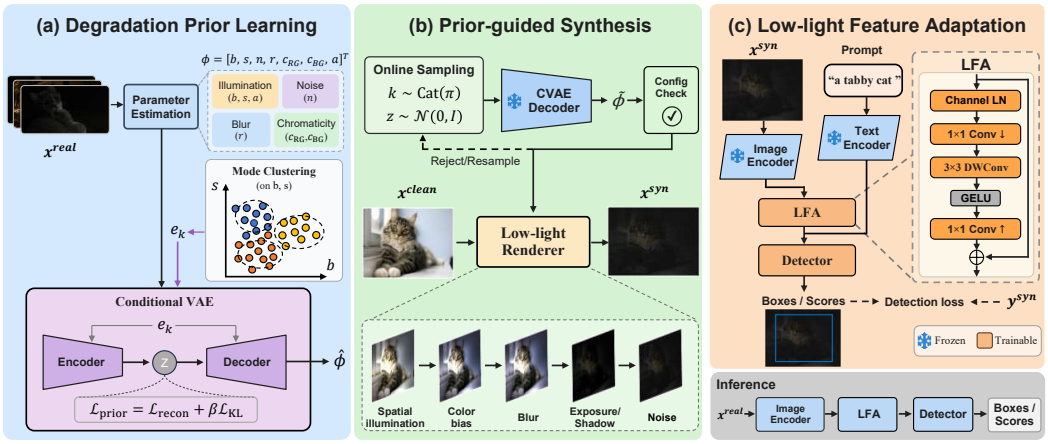


Figure 2: Overview of CoDeS. An illumination-conditioned prior learns joint degradation patterns from unlabeled target images. Its samples guide an online renderer that preserves source annotations. The LFA and detector learn from the synthesized views with frozen SAM 3 image and text encoders. At inference, real low-light images pass directly through the encoder, LFA, and detector.

3.2 Learning a Transferable Degradation Prior

Appearance descriptors. To make degradation patterns transferable across scenes, we represent each target image with a compact appearance descriptor:

$$\boldsymbol{\phi} = [b, s, n, r, c_{RG}, c_{BG}, a]^\top, \tag{1}$$

covering brightness, shadow contrast, visible noise, blur, two chromatic biases, and spatial illumination variation. These are observable appearance statistics that guide synthesis. Table 1 gives their definitions. More details are given in Appendix A.1.

Learning degradation strains. Within the prior-fitting split of $\mathcal{U}_T$, we standardize valid coordinates using reliability-weighted means $\boldsymbol{\mu}$ and standard deviations $\boldsymbol{\sigma}$, and zero-fill invalid coordinates. Confidence-weighted K-means (Lloyd, 1982) on $(\bar{b}, \bar{s})$ identifies illumination modes, such as uniformly dim scenes and scenes with pronounced shadows. We use images with valid, unclipped brightness and shadow measurements for mode discovery and CVAE fitting, and denote their fitting indices by $\mathcal{I}$. Each mode defines a degradation "strain." A single conditional variational autoencoder (CVAE) (Sohn et al., 2015) learns how all seven coordinates vary together within these strains, providing complete configurations for synthesis.

For image i, the CVAE encoder with parameters ψ produces a Gaussian posterior $q_i = q_\psi(\mathbf{z} \mid \bar{\boldsymbol{\phi}}_i, \boldsymbol{\alpha}_i, \mathbf{e}_{k_i})$, where $\boldsymbol{\alpha}_i$ contains measurement confidences and $\mathbf{e}_{k_i}$ is the one-hot mode encoding. The decoder with parameters θ jointly reconstructs the coordinates as $\widehat{\bar{\boldsymbol{\phi}}}_i = g_\theta(\mathbf{z}_i, \mathbf{e}_{k_i})$, with $\mathbf{z}_i \sim q_i$. The shared latent code models dependencies among degradation factors within each mode. We optimize ψ and θ using

$$\mathcal{L}_{\text{prior}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left[\mathbb{E}_{\mathbf{z}_i \sim q_i} \frac{\sum_{j=1}^{7} w_{ij} (\widehat{\bar{\phi}}_{ij} - \bar{\phi}_{ij})^2}{\sum_{j=1}^{7} w_{ij} + \epsilon_{\text{rec}}} + \frac{\beta}{d_z} D_{\text{KL}}(q_i \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})) \right], \tag{2}$$

where w_{ij} is the reconstruction reliability weight, d_z is the latent dimension, β controls regularization, and $\epsilon_{\text{rec}} > 0$ ensures numerical stability. Weights determine relative coordinate contributions within each sample and vanish for invalid or boundary-clipped measurements. The KL term supports subsequent sampling from a standard Gaussian. Standardization, mode selection, and prior fitting are specified in Appendices A.2.1 and A.2.2.

Table 1: Seven appearance descriptors used by CoDeS. Formulas give raw measurements before reliability weighting and bounding. References motivate the estimation principles.

Descriptor	Computation	Interpretation
Brightness (b)	$b = \log_2(Q_{0.50}(Y) + \varepsilon)$	Lower values indicate darker images.
Shadow contrast (s)	$s = \log_2 \dfrac{Q_{0.50}(Y) + \varepsilon}{Q_{0.10}(Y) + \varepsilon}$	Larger values indicate deeper shadows relative to median brightness.
Visible noise (n)	$D = Y - G_1 * Y,$ $n = 1.4826\ \mathrm{MAD}_{p \in P_{\mathrm{flat}}}\, D(p)$	Robust residual scale in flat regions; larger values indicate stronger fluctuations.
Blur (r)	$\ell = \max_{d \in \mathcal{D}_e} \left[1 - \dfrac{E_d^1}{E_d^0 + \varepsilon}\right]_+,$ $r = \left[1 + (\ell/0.15)^3\right]^{-1}$	Less edge attenuation under reblurring yields a larger blur score.
Red–green bias (c_{RG})	$c_{RG} = \log_2 \dfrac{\max(\bar{R}, \delta)}{\max(\bar{G}, \delta)}$	Positive values indicate a stronger red response relative to green.
Blue–green bias (c_{BG})	$c_{BG} = \log_2 \dfrac{\max(\bar{B}, \delta)}{\max(\bar{G}, \delta)}$	Positive values indicate a stronger blue response relative to green.
Spatial variation (a)	$z_{uv} = \log_2 \dfrac{\max(L_{uv}, \varepsilon)}{L_0},$ $a = 1.4826\ \mathrm{MAD}_{u,v}\, z_{uv}$	Robust dispersion of coarse log-brightness; larger values indicate greater nonuniformity.

Notation. For normalized sRGB channels, we use the display-domain brightness proxy $Y = 0.2126R + 0.7152G + 0.0722B$, using the coefficients from BT.709 (International Telecommunication Union, 2015). Q_τ is the τ-quantile, G_σ a normalized Gaussian kernel of standard deviation σ pixels, and $*$ convolution; $\mathrm{MAD}(V) = \mathrm{med}_{v \in V} |v - \mathrm{med}(V)|$. We use $\varepsilon = 1/255$ and $\delta = 10^{-6}$. p indexes pixels. P_{flat} contains mid-tone, low-gradient pixels excluding JPEG block boundaries. For blur, $Y_0 = G_{0.5} * Y$, $Y_1 = G_{1.2} * Y_0$, and $E_d^j = \sum_p M_e(p) |\partial_d Y_j(p)|$ ($j = 0, 1$) on a reliable-edge mask M_e. We use $\mathcal{D}_e = \{d \in \{x, y\} : E_d^0 > 0\}$ and $[v]_+ = \max(v, 0)$; an empty $\mathcal{D}_e$ gives an invalid measurement. $(\bar{R}, \bar{G}, \bar{B})$ are channel medians of linear RGB over near-neutral pixels. L_{uv} is the median of cell (u, v) in an 8×8 grid of $G_{\sigma_{\mathrm{lp}}} * Y$, with $\sigma_{\mathrm{lp}} = \max(2, \min(H, W)/16)$ and $L_0 = \max(\mathrm{med}_{u,v}\, L_{uv}, \varepsilon)$.

Basis. Robust scale (Rousseeuw & Croux, 1993); reblurring (Crété et al., 2007); diagonal channel-gain modeling (Finlayson et al., 1994).

3.3 Prior-Guided Low-Light Synthesis

After fitting the prior, we freeze the CVAE and use its decoder to sample seven-dimensional degradation configurations:

$$k \sim \mathrm{Cat}(\boldsymbol{\pi}), \qquad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad \tilde{\boldsymbol{\phi}} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot g_\theta(\mathbf{z}, \mathbf{e}_k), \tag{3}$$

where π_k is the weighted frequency of mode k. Configurations violating descriptor bounds or brightness–shadow compatibility constraints are rejected and resampled as complete vectors. We denote the accepted distribution by p_T.

For a labeled source image (x^S, y^S) and $\tilde{\boldsymbol{\phi}} \sim p_T$, an image-formation-inspired renderer generates

$$\begin{aligned} \tilde{x}^S &= \mathcal{R}(x^S; \tilde{\boldsymbol{\phi}}, \boldsymbol{\xi}) \\ &= \mathcal{J} \circ \mathcal{N}_n \circ \mathcal{T}_{b,s} \circ \mathcal{B}_r \circ \mathcal{C}_{c_{RG}, c_{BG}} \circ \mathcal{S}_a(x^S), \end{aligned} \tag{4}$$

where subscripts index the sampled coordinates and $\boldsymbol{\xi}$ specifies random illumination fields and noise realizations. Operators act from right to left, with RGB-domain conversions implicit. In linear RGB, $\mathcal{S}_a$ applies a smooth field with unit mean gain, $\mathcal{C}$ adjusts relative channel gains, and $\mathcal{B}_r$ applies Gaussian blur calibrated from r. The tone operator $\mathcal{T}_{b,s}$ then controls brightness and shadows through exposure search and a median-anchored luminance mapping, accounting for preceding appearance changes. In display RGB, $\mathcal{N}_n$ adds noise calibrated from n, including signal-independent

and luminance-dependent components (Foi et al., 2008). Quantization and JPEG encoding through $\mathcal{J}$ complete the image.

The renderer approximates the sampled statistics while preserving spatial geometry, so $(\tilde{x}^S, y^S)$ reuses the source annotations. Brightness and shadow contrast are checked after encoding; bounded retries change $\boldsymbol{\xi}$ while keeping $\tilde{\phi}$ fixed. Resampling configurations across epochs exposes each scene to different learned degradation patterns. Appendix A.3 provides calibration mappings and acceptance checks.

3.4 Low-Light Feature Adaptation

To learn from the synthesized low-light views, we insert the LFA between the frozen SAM 3 image encoder and detector. The adapter combines bottleneck channel transformations with local spatial aggregation to adapt pretrained features for object recognition and localization. For $F \in \mathbb{R}^{C\times H_f\times W_f}$,

$$A_\omega(F) = F + W_\uparrow \,\mathrm{GELU}(\mathrm{DWConv}_{3\times 3}\,(W_\downarrow \,\mathrm{CLN}(F)))\,. \tag{5}$$

Here, CLN normalizes channels at each spatial location (Ba et al., 2016). The 1×1 projections reduce and restore the channel dimension, while the depth-wise convolution preserves spatial dimensions and aggregates local context before GELU (Hendrycks & Gimpel, 2016). The bottleneck limits additional parameters, and the residual connection retains the pretrained feature path. The LFA takes only image features as input.

Training and inference. After prior fitting, we jointly train the LFA and detector using source detection labels, with the prior and both SAM 3 encoders fixed. Let $\mathcal{D}_{\mathrm{syn}}$ denote the distribution of synthesized source pairs admitted by the rendering checks. With image encoder E_v, text encoder E_t, category prompts t, and detector H_ν, we optimize

$$\min_{\omega,\nu} \quad \mathbb{E}_{(\tilde{x}^S,y^S)\sim\mathcal{D}_{\mathrm{syn}}}\left[\mathcal{L}_{\mathrm{det}}\big(H_\nu(A_\omega(E_v(\tilde{x}^S)), E_t(t)), y^S\big)\right], \tag{6}$$

where $\mathcal{L}_{\mathrm{det}}$ is the base detector's detection loss. Only LFA parameters ω and detector parameters ν are updated. At inference, a real low-light image x^T is processed as $H_\nu(A_\omega(E_v(x^T)), E_t(t))$ to predict boxes and category scores. The LFA remains active; descriptor extraction, prior sampling, and rendering are confined to training.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation. We evaluate on three real low-light datasets: ExDark (Loh & Chan, 2019), NOD (Morawski et al., 2021), and DARK FACE (Yang et al., 2020).ExDark covers 12 categories with an official split of 3,000/1,800/2,563 training/validation/test images. NOD combines Sony and Nikon RGB images for three categories and uses the official 5,777/721/721 split. DARK FACE uses 3,600/1,200/1,200 images, following FRBNet (Sun et al., 2025). Each task uses 3,000 labeled normal-light source images: COCO (Lin et al., 2014) for ExDark and NOD, retaining the corresponding target categories, and WIDER FACE (Yang et al., 2016) for DARK FACE. Source and target images are unpaired. We report Recall and mAP50.

Model and training. For each target dataset, we fit a separate illumination-conditioned CVAE to seven-dimensional degradation descriptors from unlabeled target training images, then freeze it to synthesize low-light views of labeled source images. We initialize SAM 3 (Carion et al., 2026) from official pretrained weights, freeze its image and text encoders, and jointly train the detection module and LFA using source annotations and the SAM 3 detection loss. Target validation labels are used for checkpoint and configuration selection. Each run uses a single NVIDIA RTX 3090 GPU. At inference, only SAM 3 and the LFA are used.Configuration details are provided in Appendix A.6.

4.2 Low-Light Object Detection Results

Quantitative comparison. Table 2 compares pretrained SAM 3 with supervised low-light methods, and domain adaptation methods, including Adaptive Teacher (Li et al., 2022). CoDeS uses

Table 2: Low-light object detection performance on ExDark, NOD, and DARK FACE (%).

Method	ExDark		NOD		DARK FACE	
	Recall	mAP50	Recall	mAP50	Recall	mAP50
SAM 3 (Direct inference)	88.0	77.7	66.8	60.6	50.2	47.4
(a) Supervised						
IAT [BMVC 2022]	85.0	72.6	64.3	71.2	73.6	55.5
FeatEnHancer [ICCV 2023]	90.4	71.2	61.8	68.7	74.1	55.2
YOLA [NeurIPS 2024]	86.1	72.7	63.2	69.4	74.9	56.3
FRBNet [NeurIPS 2025]	90.6	74.9	63.5	69.9	**75.7**	57.7
YES [CVIU 2025]	86.5	83.7	83.4	76.7	74.9	57.6
Ours (Real)	89.2	86.7	83.3	76.6	72.6	57.9
Ours (Real + Synthetic)	**91.7**	**88.0**	**83.9**	**77.6**	73.4	**58.0**
(b) Domain adaptation						
Adaptive Teacher [CVPR 2022]	59.7	50.9	53.6	52.3	25.4	24.9
2PCNet [CVPR 2023]	65.0	43.8	62.7	54.4	40.7	28.2
ISP-Teacher [AAAI 2024]	44.1	39.5	49.3	46.5	25.2	24.6
Ours	**92.2**	**80.9**	**71.0**	**63.4**	**59.1**	**50.5**

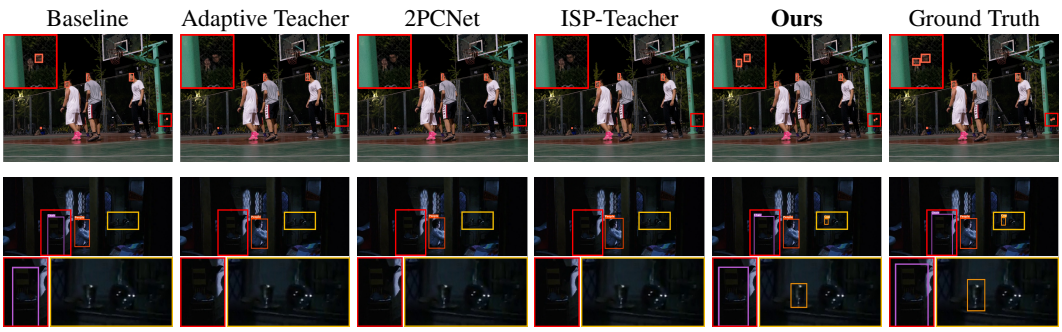


Figure 3: Qualitative comparison of detection results on two low-light scenes.

source detection annotations and fits its degradation prior on unlabeled target training images, whereas supervised methods use target training labels. The ExDark and DARK FACE entries for IAT (Cui et al., 2022), FeatEnHancer, YOLA, and FRBNet are the YOLOv3 results reported by Sun et al. (2025). It can be seen that our Real+Synthetic variant achieves the highest supervised mAP50 on ExDark, NOD, and DARK FACE, reaching 88.0%, 77.6%, and 58.0%, respectively. These results outperform the strongest supervised baselines by clear margins. Compared with our Real-only variant, synthetic augmentation contributes gains of 1.3, 1.0, and 0.1 points, indicating that the generated views can complement real low-light supervision. In the domain adaptation group, CoDeS ranks first in both mAP50 and Recall across all three benchmarks, outperforming the strongest competing methods in mAP50 by 30.0, 9.0, and 22.3 points, respectively. Compared to pretrained SAM 3, it improves mAP50 by 3.2, 2.8 and 3.1 points, with the largest Recall gain of 8.9 points on DARK FACE. These improvements support the effectiveness of combining target-informed synthesis with feature adaptation to transfer source supervision to real low-light images.

Qualitative comparison. Figure 3 compares CoDeS with pretrained SAM 3 and three domain adaptation methods on two challenging low-light scenes. In the first scene, the enlarged region contains tiny faces with limited visual detail. CoDeS detects both faces, while the baseline misses one and the compared adaptation methods leave them undetected. In the second scene, CoDeS recovers the cup that is barely visible in the dark background and missed by all compared methods. Its predicted chair box also more closely matches the ground-truth extent. These examples illustrate improved detection of small and poorly illuminated objects, suggesting that target-informed synthesis and feature adaptation help the detector recognize objects when visual cues are weak.

4.3 Synthetic Low-Light Images

We assess synthesized images through distributional similarity on DARK FACE and downstream detection performance on ExDark.

Table 3: Distributional similarity between synthesized low-light images and real DARK FACE images ($N = 1200$). KID and Lab-SWD values are multiplied by 10^3. $W_1(Y)$ measures the Wasserstein distance between image-balanced luminance histograms. Best and second-best results are bold and underlined, respectively.

Synthesis method	FID ↓	KID ↓	Lab-SWD ↓	$W_1(Y)$ ↓
Crafting [BMVC 2021]	248.05	188.76	93.36	0.17
Similarity Min-Max [ICCV 2023]	<u>213.75</u>	**141.22**	70.70	0.13
DEN [McGE 2025]	227.98	164.94	<u>20.03</u>	<u>0.05</u>
ELVIS [CVPR 2026]	242.83	177.48	236.30	0.49
CoDeS [Ours]	**212.27**	<u>149.65</u>	**10.33**	**0.02**

Distributional similarity. CoDeS achieves the lowest FID, Lab-SWD, and $W_1(Y)$ among the compared methods (Table 3). Relative to DEN, the strongest alternative on Lab-SWD and $W_1(Y)$, it reduces these distances by 48.4% and 60.0%, respectively, indicating a closer match to real low-light images in local appearance and luminance. Its advantage in Inception-feature space is less pronounced: FID improves slightly over Similarity Min-Max (212.27 versus 213.75), while Similarity Min-Max achieves the lowest KID. The strongest gains therefore concern the appearance and luminance statistics that motivate our degradation modeling. Figure 4 provides a visual example: CoDeS produces a dark background and partially illuminated foreground objects, introducing spatial variation in illumination within the same scene.

Utility for detection. We further assess the generated images by training SAM 3 with the LFA using each synthesis method and evaluating on ExDark (Table 4). CoDeS achieves 80.9% mAP50, outperforming all competing methods and improving over pretrained SAM 3 by 3.2 points. Because the synthesis variants share the detector and adapter architectures, the advantage over competing generators supports the value of CoDeS-generated training views for low-light adaptation. Together with the distributional comparison, these results show that CoDeS improves both the match to measured target appearance and the usefulness of synthetic images for detection.

Table 4: Detection on ExDark with different synthesis methods, all using SAM 3+LFA.

Synthesis method	mAP50 (%)	Gain (pp)
Baseline (SAM 3)	77.7	—
Crafting	78.9	+1.2
Similarity Min-Max	79.4	+1.7
DEN	79.4	+1.7
ELVIS	79.2	+1.5
Ours	**80.9**	**+3.2**

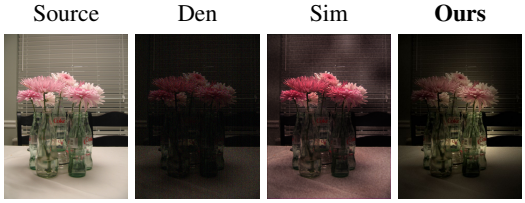


Figure 4: Low-light images synthesized from the same source scene by different methods. Sim: Similarity Min-Max.

4.4 Ablation Studies

Table 5: Ablation of degradation sampling and LFA on ExDark.

Method	**Degradation Strategy**	**LFA**	**mAP50 (%)**
SAM 3	—	✗	77.7
	Empirical Sampling	✗	79.1
	Empirical Sampling	✓	80.0
CoDeS	CVAE Generation	✗	79.6
	CVAE Generation	✓	**80.9**

Table 6: Effect of illumination modes K on mAP50 (%).

K	ExDark	NOD	DARK FACE
1	79.4	61.4	50.0
2	79.8	61.8	48.0
3	**80.9**	**63.4**	**50.5**
4	79.2	61.7	48.0
5	78.3	60.9	48.6

Component contributions. Table 5 compares two degradation strategies with and without the LFA on ExDark. *Empirical Sampling* randomly draws a complete seven-dimensional degradation vector from the bank, with all parameters taken from the same unlabeled target training image. *CVAE Generation* samples complete configurations from the learned illumination-conditioned

CVAE, modeling the joint variation of degradation parameters. Both strategies use the same renderer. Without the LFA, empirical sampling and CVAE generation achieve 79.1% and 79.6% mAP50, respectively, compared with 77.7% for pretrained SAM 3. Adding the LFA raises their scores to 80.0% and 80.9%, providing further gains of 0.9 and 1.3 percentage points.The CVAE outperforms empirical sampling by 0.5 points without LFA and by 0.9 points with it, indicating that learning a conditional degradation distribution provides a more effective sampling strategy than directly drawing from the empirical parameter bank. These results show that reusing measured target configurations already provides useful training variation, while sampling from a learned conditional prior offers an additional improvement. The LFA benefits both sampling strategies.

Number of illumination modes. Table 6 shows that $K = 3$ achieves the highest mAP50 on all three datasets. Compared with a single mode, three modes improve mAP50 by 1.5, 2.0, and 0.5 percentage points on ExDark, NOD, and DARK FACE, respectively. Increasing K to four or five lowers accuracy relative to $K = 3$. These results support representing distinct illumination regimes when learning the degradation prior, while showing that finer partitions do not necessarily yield more useful synthesis conditions.

4.5 Analysis of Illumination Priors and Feature Adaptation

Learned illumination modes. Figure 5 examines the prior in the brightness–shadow-gap plane. C1 has a larger shadow gap, C2 is relatively brighter with a smaller shadow gap, and C3 captures darker scenes. Generated contours broadly follow the locations and shapes of the real modes, with differences in spread. The prior thus retains distinct illumination regimes in this projection, complementing the detection gains from using multiple modes. This two-dimensional view does not establish a match of the full degradation distribution.

Feature changes induced by the LFA. Figure 6 visualizes residual update magnitudes and feature-direction changes. The latter are smaller over much of the annotated chair and larger in parts of the background, showing spatially selective feature updates. These maps illustrate the LFA's behavior on one image; its detection benefit is supported by the gains under both sampling strategies in Table 5.

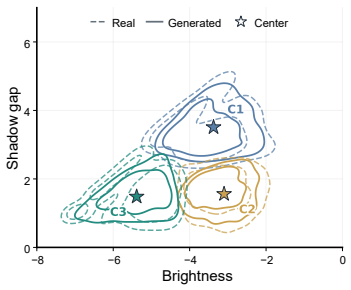
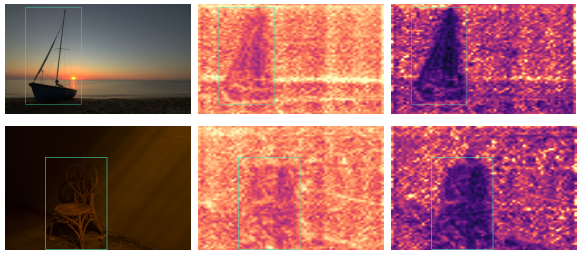


Figure 5: Real (dashed) and generated (solid) distributions at $K = 3$. Stars mark cluster centers.



(a) Input (b) RMS (c) Direction

Figure 6: LFA feature changes: (a) input, (b) RMS residual magnitude, and (c) direction change. Green boxes mark ground truth.

5 Conclusion

We have presented CoDeS, which turns unlabeled low-light images into reusable sources of degradation patterns for detector adaptation. Its illumination-conditioned CVAE learns the joint variation of appearance descriptors, enabling an online renderer to transfer these patterns across labeled normal-light scenes. The LFA connects the resulting training variation to detection by adapting frozen SAM 3 features using source annotations. Experiments on three benchmarks demonstrate consistent detection gains, while synthesis comparisons support both the similarity of generated images to measured target appearance and their value for detector training. These findings show that learning degradation patterns from existing target images can expand low-light training data while reusing source supervision.

AI USE STATEMENT

Generative AI tools were used in this work to assist with language translation and polishing, literature search and summarization, and the refinement of manuscript presentation. All AI-assisted content was reviewed and revised by the authors, and factual claims were verified against the original sources. The research methodology, experimental design, analysis, conclusions, and final manuscript content were determined by the authors, who take full responsibility for the work.

REPRODUCIBILITY STATEMENT

The CoDeS formulation, LFA architecture, and training and inference procedures are described in Section 3. Appendices A.1, A.2, and A.3 detail descriptor extraction and preprocessing, illumination-mode discovery and CVAE fitting, and rendering controls, calibration, and acceptance checks. Section 4.1 specifies the source and target datasets, data splits, evaluation metrics, and supervision protocol. Appendix A.6 reports the prior configurations, optimization hyperparameters, image preprocessing, random seed, hardware, and validation-based checkpoint selection used in our experiments.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026.

Frédérique Crété, Thierry Dolmière, Patricia Ladret, and Marina Nicolas. The blur effect: Perception and estimation with a new no-reference perceptual blur metric. In *Human Vision and Electronic Imaging XII*, volume 6492, pp. 64920I, 2007. doi: 10.1117/12.702790.

Yiming Cui, Liang Li, Haibing Yin, Yuhan Gao, Yaoqi Sun, and Chenggang Yan. Debiased teacher for day-to-night domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2577–2587, 2025.

Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask AET with orthogonal tangent regularity for dark object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2553–2562, 2021.

Ziteng Cui, Kunchang Li, Lin Gu, Shenghan Su, Peng Gao, Zhengkai Jiang, Yu Qiao, and Tatsuya Harada. You only need 90K parameters to adapt light: A light weight transformer for image enhancement and exposure correction. In *Proceedings of the British Machine Vision Conference*, 2022.

Zhipeng Du, Miaojing Shi, and Jiankang Deng. Boosting object detection with zero-shot day-night domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12666–12676, 2024.

James H. Elder and Steven W. Zucker. Local scale control for edge detection and blur estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(7):699–716, 1998. doi: 10.1109/34.689301.

Graham D. Finlayson, Mark S. Drew, and Brian V. Funt. Color constancy: Generalized diagonal transforms suffice. *Journal of the Optical Society of America A*, 11(11):3011–3019, 1994. doi: 10.1364/JOSAA.11.003011.

Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian–gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008. doi: 10.1109/TIP.2008.2001399.

Khurram Azeem Hashmi, Goutham Kallempudi, Didier Stricker, and Muhammad Zeshan Afzal. FeatEnHancer: Enhancing hierarchical features for object detection and beyond under low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6725–6735, 2023.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs), 2016.

Mingbo Hong, Shen Cheng, Haibin Huang, Haoqiang Fan, and Shuaicheng Liu. You only look around: Learning illumination-invariant feature for low-light object detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87136–87158, 2024. doi: 10.52202/079017-2765.

Yang Hong, Kaixuan Wei, Linwei Chen, and Ying Fu. Crafting object detection in very low light. In *Proceedings of the British Machine Vision Conference*, 2021. doi: 10.5244/C.35.21.

Jiayi Hu, Qian Zhao, and Gang Li. DLDA: Unified dual-level domain adaptation for low-light object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4851–4859, 2026. doi: 10.1609/aaai.v40i6.42488.

International Telecommunication Union. Parameter values for the HDTV standards for production and international programme exchange. Recommendation ITU-R BT.709-6, International Telecommunication Union, June 2015.

Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T. Tan. 2PCNet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11484–11493, 2023.

Edwin H. Land and John J. McCann. Lightness and Retinex theory. *Journal of the Optical Society of America*, 61(1):1–11, 1971. doi: 10.1364/JOSA.61.000001.

Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7581–7590, 2022.

Joanne Lin, Crispian Morris, Ruirui Lin, Fan Zhang, David Bull, and Nantheera Anantrasirichai. Towards a general-purpose zero-shot synthetic low-light image and video pipeline. In *Proceedings of the 3rd International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice*, pp. 3–11, 2025. doi: 10.1145/3746278.3759376.

Joanne Lin, Ruirui Lin, Yini Li, David Bull, and Nantheera Anantrasirichai. ELVIS: Enhance low-light for video instance segmentation in the dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25011–25021, 2026.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. doi: 10.1016/j.cviu.2018.10.010.

Rundong Luo, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Similarity min-max: Zero-shot day-night domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8104–8114, 2023.

Igor Morawski, Yu-An Chen, Yu-Sheng Lin, and Winston H. Hsu. NOD: Taking a closer look at detection under extreme low-light conditions with night object detection dataset. In *Proceedings of the British Machine Vision Conference*, 2021. doi: 10.5244/C.35.317.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

Peter J. Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993. doi: 10.1080/01621459.1993.10476408.

Gwangik Shin, Jaeha Song, and Soonmin Hwang. Self-guided low light object detection framework. In *The Fourteenth International Conference on Learning Representations*, 2026.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28, pp. 3483–3491, 2015.

Fangtong Sun, Congyu Li, Ke Yang, Yuchen Pan, Hanwen Yu, Xichuan Zhang, and Yiying Li. FRBNet: Revisiting low-light vision through frequency-domain radial basis network. In *Advances in Neural Information Processing Systems*, volume 38, pp. 99085–99113, 2025. doi: 10.52202/085713-3309.

Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. WIDER FACE: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5525–5533, 2016.

Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J. Scheirer, Zhangyang Wang, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020. doi: 10.1109/TIP.2020.2981922.

Shu Ye, Wenxin Huang, Wenxuan Liu, Liang Chen, Xiao Wang, and Xian Zhong. YES: You should examine suspect cues for low-light object detection. *Computer Vision and Image Understanding*, 251:104271, 2025. doi: 10.1016/j.cviu.2024.104271.

Yin Zhang, Yongqiang Zhang, Zian Zhang, Man Zhang, Rui Tian, and Mingli Ding. ISP-Teacher: Image signal process with disentanglement regularization for unsupervised domain adaptive dark object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7387–7395, 2024. doi: 10.1609/aaai.v38i7.28569.

A Appendix

Appendix Contents

A.1 Degradation Statistics 14
A.1.1 Analysis Scale, Notation, and Descriptor Ranges 14
A.1.2 Brightness and Shadow Contrast 14
A.1.3 Visible-Noise Estimation 15
A.1.4 Blur Estimation by Relative Edge Attenuation 15
A.1.5 Relative Chromatic Bias 16
A.1.6 Spatial Illumination Variation 17
A.2 Prior Preparation and CVAE Training 17
A.2.1 Bounds, Missing Observations, and Standardization 17
A.2.2 Illumination-Mode Discovery and Prior Fitting 18
A.3 Rendering Controls and Calibration 19
A.3.1 Accepted Configurations and Rendering Targets 19
A.3.2 Spatial and Chromatic Controls 19
A.3.3 Blur and Noise Calibration 20
A.3.4 Exposure–Shadow Mapping 20
A.3.5 Terminal Checks and Failure Handling 21
A.4 Ablation of Degradation Dimensions 22
A.5 Effectiveness of Synthesis Across Detectors 23
A.6 Experimental Configuration 23

A.1 Degradation Statistics

This appendix specifies the seven appearance descriptors used in Section 3.2, including their measurement validity and confidence estimates. The descriptors summarize target appearance to guide synthesis; their raw definitions correspond to Table 1. Prior preparation and CVAE training are described separately in Appendix A.2. These statistics can also reflect scene content and camera processing, so they are observable appearance coordinates rather than calibrated sensor parameters.

A.1.1 Analysis Scale, Notation, and Descriptor Ranges

We preserve aspect ratio and downsample images whose longer side exceeds 512 pixels. Throughout this section, H and W refer to this analysis resolution. For sRGB values $x \in [0,1]^{H\times W\times 3}$, the display-domain brightness proxy is

$$Y(p) = 0.2126R(p) + 0.7152G(p) + 0.0722B(p), \qquad \varepsilon = 1/255, \quad \delta = 10^{-6}. \tag{7}$$

The coefficients in Eq. (7) follow BT.709 (International Telecommunication Union, 2015); applied to sRGB values here, they define a display-domain brightness proxy. Here $Q_\tau(V)$ is the τ-quantile of values in V, med denotes the median, and $\mathrm{MAD}(V) = \mathrm{med}_{v\in V}\,|v - \mathrm{med}(V)|$. We write $\mathrm{clip}_{[l,u]}(v) = \min\{u, \max\{l, v\}\}$, $[v]_+ = \max\{v, 0\}$, and use $\mathbb{I}[\cdot]$ for an indicator. Coordinates follow the order $\boldsymbol{\phi} = [b, s, n, r, c_{RG}, c_{BG}, a]^\top$ in Eq. (1). Each has a raw estimate $\phi_{j,\mathrm{raw}}$, a bounded value ϕ_j, a validity flag $m_j \in \{0,1\}$, and confidence $\alpha_j \in [0,1]$. The chromatic coordinates c_{RG}, c_{BG} are distinct from the measurement confidences collected in $\boldsymbol{\alpha}$. We define the coordinate-wise working ranges in Table 7 to bound the appearance configurations considered during prior fitting and rendering. The blur interval $r \in [0,1]$ follows from its normalized score, while the other intervals are prescribed design choices.

The confidence formulas below apply when their supporting statistics are defined and the corresponding validity conditions hold. Otherwise, we set $m_j = 0$ and $\alpha_j = 0$ and omit that coordinate from measurement-based sums. Undefined statistics, including medians of empty sets, are not evaluated as observations. Invalid coordinates are encoded only after standardization, as specified in Appendix A.2.1.

Table 7: Observable degradation coordinates and their prescribed ranges.

Coordinate	Range	Observable quantity
b	$[-8, -0.5]$	Log-median display luma
s	$[0, 7]$	Median-to-shadow log-quantile gap
n	$[0, 0.08]$	Robust residual noise scale
r	$[0, 1]$	Relative edge response to reblurring
c_{RG}	$[-1, 1]$	Linear-RGB red-to-green log ratio
c_{BG}	$[-1, 1]$	Linear-RGB blue-to-green log ratio
a	$[0, 4]$	Low-frequency spatial luma dispersion

A.1.2 Brightness and Shadow Contrast

We define global brightness and shadow contrast by

$$b_{\mathrm{raw}} = \log_2(Q_{0.50}(Y) + \varepsilon), \qquad s_{\mathrm{raw}} = \log_2 \frac{Q_{0.50}(Y) + \varepsilon}{Q_{0.10}(Y) + \varepsilon}. \tag{8}$$

The first quantity decreases with overall image darkness; the second describes the relative separation between mid-tones and the lower luminance tail. The logarithmic representation provides a compact description of multiplicative intensity changes, consistent with the use of relative intensities in illumination modeling (Land & McCann, 1971). Since these statistics are evaluated in sRGB, we interpret them as display-appearance coordinates rather than physical exposure measurements.

Brightness has $m_b = \mathbb{I}[Q_{0.50}(Y) > 0]$, with $\alpha_b = 1$ for valid observations. For shadows, let $d_Y = Q_{0.90}(Y) - Q_{0.10}(Y)$ and set

$$m_s = \mathbb{I}[d_Y \geq 4/255], \qquad \alpha_s = \mathrm{clip}_{[0,1]}(d_Y/0.20). \tag{9}$$

This reduces the influence of nearly constant or black images, in which quantization can dominate the estimated shadow gap.

A.1.3 VISIBLE-NOISE ESTIMATION

We estimate residual noise in relatively flat, mid-tone regions. Let G_σ be a normalized isotropic Gaussian kernel with standard deviation σ (in pixels), and let S_x and S_y denote the Sobel derivative operators. We define

$$Y_s = G_1 * Y, \qquad D = Y - Y_s, \qquad g_{\text{flat}}(p) = \sqrt{(S_x * Y_s)^2(p) + (S_y * Y_s)^2(p)}. \tag{10}$$

Candidate pixels satisfy a mid-tone condition and a weak-gradient condition:

$$\begin{aligned} M_{\text{mid}}(p) &= \mathbb{I}[Q_{0.10}(Y) \le Y(p) \le Q_{0.90}(Y)], \\ M_{\text{flat}}(p) &= M_{\text{mid}}(p)\, \mathbb{I}[g_{\text{flat}}(p) \le Q_{0.35}(g_{\text{flat}})]\, \mathbb{I}[p \notin \mathcal{B}_8]. \end{aligned} \tag{11}$$

The set $\mathcal{B}_8$ denotes exclusion bands around an 8×8-pixel block grid at the analysis resolution. This exclusion reduces sensitivity to grid-aligned artifacts; after resizing, it need not coincide with the original JPEG block boundaries. On a nonempty $P_{\text{flat}} = \{p : M_{\text{flat}}(p) = 1\}$, we compute

$$n_{\text{raw}} = 1.4826 \ \text{med}_{p\in P_{\text{flat}}} \left| D(p) - \text{med}_{q\in P_{\text{flat}}} D(q) \right|. \tag{12}$$

The factor 1.4826 gives Gaussian-consistent scaling of the MAD (Rousseeuw & Croux, 1993). The result measures filtered residual variation in the displayed image, not an independently calibrated sensor standard deviation. This distinction is relevant because signal-dependent acquisition noise is altered by the camera pipeline (Foi et al., 2008).

For a collection $\mathcal{G}$ of local regions containing sufficient flat pixels, we also compute regional estimates n_Ω using the same rule. Confidence combines the number of supporting pixels and agreement between regions:

$$\alpha_n = \min\left\{1, \frac{|P_{\text{flat}}|}{4096}\right\} \exp\left(-\frac{\text{std}_{\Omega\in\mathcal{G}}(n_\Omega)}{|\mathcal{G}|^{-1}\sum_{\Omega\in\mathcal{G}} n_\Omega + \varepsilon}\right). \tag{13}$$

We set $m_n = 1$ only when $|P_{\text{flat}}| \ge 256$ and $|\mathcal{G}| \ge 2$; otherwise $\alpha_n = 0$ and the coordinate is excluded from normalization and reconstruction supervision.

A.1.4 BLUR ESTIMATION BY RELATIVE EDGE ATTENUATION

We compare edge responses before and after additional smoothing. Reblurring-based measures exploit the smaller visible change produced by smoothing an already blurred image (Crété et al., 2007); edge-scale analysis provides related context for contour-based blur estimation (Elder & Zucker, 1998). We first compute $Y_0 = G_{0.5} * Y$, its horizontal and vertical gradients (g_x, g_y), and the structure tensor

$$J = G_{1.2} * \begin{bmatrix} g_x^2 & g_x g_y \\ g_x g_y & g_y^2 \end{bmatrix}, \qquad \kappa(p) = \frac{\sqrt{(J_{xx} - J_{yy})^2 + 4J_{xy}^2}}{J_{xx} + J_{yy} + 10^{-8}}. \tag{14}$$

Here J averages local gradient products, and $\kappa \in [0,1]$ measures directional coherence: values near one indicate a dominant edge direction, while smaller values indicate weak or mixed directional structure. The constant 10^{-8} stabilizes the denominator. With gradient magnitude $g_{\text{edge}} = \sqrt{g_x^2 + g_y^2}$, we retain strong, coherent edges using

$$M_e(p) = \mathbb{I}[g_{\text{edge}}(p) \ge \max\{Q_{0.85}(g_{\text{edge}}), 0.01\} \ \wedge \ \kappa(p) \ge 0.60]. \tag{15}$$

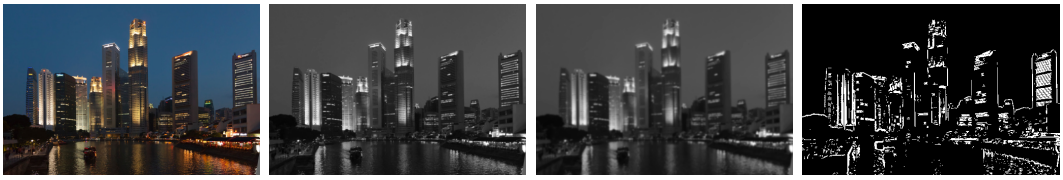
The quantile threshold selects relatively strong gradients, the floor 0.01 excludes weak responses, and $\kappa \ge 0.60$ requires directional coherence. Figure 7 illustrates the two smoothing stages and the selected edge support. We then form $Y_1 = G_{1.2} * Y_0$ and measure edge attenuation on the same mask, so that the comparison uses identical pixel locations:

$$\begin{aligned} E_d^j &= \sum_p M_e(p) |\partial_d Y_j(p)|, \qquad d \in \{x,y\}, \quad j \in \{0,1\}, \\ \mathcal{D}_e &= \{d \in \{x,y\} : E_d^0 > 0\}, \\ \ell &= \max_{d\in\mathcal{D}_e} \left[1 - \frac{E_d^1}{E_d^0 + \varepsilon}\right]_+, \qquad r_{\text{raw}} = \left[1 + (\ell/0.15)^3\right]^{-1}. \end{aligned} \tag{16}$$

Here ∂_d is the spatial derivative in direction d, and E_d^0, E_d^1 sum the absolute responses before and after reblurring. Their ratio measures the retained edge response; ℓ takes the largest nonnegative fractional loss over supported directions. Directions with $E_d^0 = 0$ are excluded because they provide no attenuation evidence; an empty $\mathcal{D}_e$ makes the measurement invalid. Small ℓ indicates little additional attenuation and thus a larger blur score. We define the decreasing mapping with half-response $r_{\text{raw}} = 0.5$ at $\ell = 0.15$ and exponent 3 controlling its steepness. Its conversion to Gaussian rendering width is specified in Appendix A.3. With $N_e = \sum_p M_e(p)$ denoting the number of selected edge pixels, validity and confidence are

$$m_r = \mathbb{I}[N_e \geq 32 \ \wedge \ \mathcal{D}_e \neq \varnothing], \qquad \alpha_r = 0.60 \min\left\{1, \frac{N_e}{\max\{0.02HW, 64\}}\right\} \text{med}_{p:M_e(p)=1}\, \kappa(p). \tag{17}$$

Validity requires at least 32 supporting pixels. Confidence combines edge coverage, normalized by $\max\{0.02HW, 64\}$, with median directional coherence; the factor 0.60 caps confidence at 0.60. These thresholds and mapping constants are fixed design choices. Invalid measurements have $\alpha_r = 0$, and confidence is evaluated only when its supporting statistics are defined.



(a) Observed image (b) Prefiltered Y_0 (c) Reblurred Y_1 (d) Edge mask M_e

Figure 7: Edge selection for the reblurring statistic. (a) Observed RGB image. (b) Luma after mild prefiltering, $Y_0 = G_{0.5} * Y$. (c) Additional smoothing, $Y_1 = G_{1.2} * Y_0$, softens window and building edges. (d) White pixels mark strong, directionally coherent edges selected from Y_0 by Eq. (15). Responses in (b) and (c) are compared at these same locations; the mask is not recomputed after reblurring.

A.1.5 Relative Chromatic Bias

Chromatic descriptors use linear RGB. Each sRGB channel v is transformed by

$$\Gamma^{-1}(v) = \begin{cases} v/12.92, & v \leq 0.04045, \\ ((v + 0.055)/1.055)^{2.4}, & v > 0.04045. \end{cases} \tag{18}$$

We select candidate pixels whose display luma lies in $[\max\{Q_{0.10}(Y), 2/255\}, Q_{0.95}(Y)]$; Figure 8 illustrates the spatial selection on an example image. Their saturation is measured before linearization over channels $\chi \in \{R, G, B\}$ as

$$\text{sat}(p) = \frac{\max_\chi x_\chi(p) - \min_\chi x_\chi(p)}{\max_\chi x_\chi(p) + \delta}. \tag{19}$$

If $N_{\text{cand}} < 128$, both color estimates are invalid. Otherwise, we take the $K_{\text{col}} = \max\{64, \lceil 0.25 N_{\text{cand}} \rceil\}$ least saturated candidates as a near-neutral set P_{neutral}. Relative channel responses are motivated by diagonal channel-gain models (Finlayson et al., 1994); the neutral-pixel selection reduces interference from strongly colored surfaces. It does not guarantee recovery of the scene illuminant.



(a) Observed image (b) Selected candidates

Figure 8: Pixel selection for color statistics. The cyan overlay in (b) shows selected candidate regions in the image in (a). Color estimation uses low-saturation pixels within the admissible luma range.

Let $(\bar{R}, \bar{G}, \bar{B})$ be the channel-wise medians of the linear RGB values in P_{neutral}. The descriptors are

$$c_{RG,\text{raw}} = \log_2 \frac{\max\{\bar{R}, \delta\}}{\max\{\bar{G}, \delta\}}, \qquad c_{BG,\text{raw}} = \log_2 \frac{\max\{\bar{B}, \delta\}}{\max\{\bar{G}, \delta\}}. \tag{20}$$

Let $\overline{\mathrm{sat}}$ and $\bar{Y}_{P_{\mathrm{neutral}}}$ denote median saturation and display luma in P_{neutral}. The two coordinates share confidence

$$\alpha_{\mathrm{col}} = \min\left\{1, \frac{K_{\mathrm{col}}}{2048}\right\}\left[1 - \mathrm{clip}_{[0,1]}(\overline{\mathrm{sat}}/0.35)\right]\frac{\bar{Y}_{P_{\mathrm{neutral}}}}{\bar{Y}_{P_{\mathrm{neutral}}} + 0.02}. \tag{21}$$

Both are valid only if $N_{\mathrm{cand}} \geq 128$, $|P_{\mathrm{neutral}}| \geq 64$, and $\overline{\mathrm{sat}} \leq 0.35$. For these observations, the two corresponding entries of $\boldsymbol{\alpha}$ equal α_{col}; otherwise both are zero.

A.1.6 Spatial Illumination Variation

We first suppress high-frequency structure using

$$Y_{\mathrm{lp}} = G_{\sigma_{\mathrm{lp}}} * Y, \qquad \sigma_{\mathrm{lp}} = \max\{2, \min(H, W)/16\}. \tag{22}$$

Divide Y_{lp} into an 8×8 grid and let L_{uv} denote the median in cell (u, v). Relative log-luma values are

$$L_0 = \max\{\mathrm{med}_{u,v}\, L_{uv}, \varepsilon\}, \qquad z_{uv} = \log_2 \frac{\max\{L_{uv}, \varepsilon\}}{L_0}. \tag{23}$$

The descriptor and its reliability are

$$a_{\mathrm{raw}} = 1.4826\ \mathrm{MAD}_{u,v}(z_{uv}), \qquad m_a = \mathbb{I}[L_0 \geq 2/255], \qquad \alpha_a = 0.60\frac{L_0}{L_0 + 0.03}. \tag{24}$$

Normalizing by a global reference reduces sensitivity to a uniform multiplicative luma change when the numerical floor is inactive. Large-scale scene reflectance can still influence this appearance descriptor.

A.2 Prior Preparation and CVAE Training

This appendix describes how the measurements in Appendix A.1 are prepared for learning the illumination-conditioned degradation prior in Section 3.2. We specify reliability weighting, missing-observation handling, standardization, illumination-mode discovery, and CVAE fitting and holdout evaluation.

A.2.1 Bounds, Missing Observations, and Standardization

For a defined measurement, the recorded bounded coordinate is $\phi_{ij} = \mathrm{clip}_{[l_j,u_j]}(\phi_{ij,\mathrm{raw}})$, using Table 7. Here i indexes images and j indexes the seven coordinates. We retain raw values, validity flags, confidences, and failure reasons. A clipped value indicates an out-of-range measurement rather than an exact observation at the boundary. For equally weighted images, the reconstruction reliability weights in Eq. (2) are

$$w_{ij} = \begin{cases} \alpha_{ij}, & m_{ij} = 1 \text{ and } \phi_{ij,\mathrm{raw}} = \phi_{ij}, \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

Thus invalid and clipped coordinates provide no pointwise reconstruction supervision. Within each sample, the normalization by $\sum_j w_{ij} + \epsilon_{\mathrm{rec}}$ in Eq. (2) makes these weights primarily control the relative contributions of the observed coordinates; it does not apply an additional image-level confidence weight.

Before fitting, we separate a grouped holdout from the designated target prior-training data. All normalization statistics below use only the prior-fitting split, and each sum omits zero-weight coordinates. Provided that every coordinate has positive total weight, its weighted mean and standard deviation are

$$\begin{aligned} \mu_j &= \frac{\sum_i w_{ij}\phi_{ij}}{\sum_i w_{ij}}, \\ \sigma_j &= \max\left\{\sqrt{\frac{\sum_i w_{ij}(\phi_{ij} - \mu_j)^2}{\sum_i w_{ij}}}, \delta\right\}. \end{aligned} \tag{26}$$

A coordinate with no positive-weight observations cannot be fitted by this procedure; the variance floor does not supply missing evidence. For valid coordinates, $\bar{\phi}_{ij} = (\phi_{ij} - \mu_j)/\sigma_j$; invalid coordinates are filled with zero and have $\alpha_{ij} = 0$. This fill value encodes a missing input, not an

observation of zero degradation. A valid but clipped coordinate retains its bounded input value and measurement confidence, while its reconstruction weight remains zero. The encoder receives $(\bar{\boldsymbol{\phi}}_i, \boldsymbol{\alpha}_i, \mathbf{e}_{k_i})$, as in Section 3.2; validity is handled through input encoding and the weights rather than an additional encoder input.

A.2.2 Illumination-Mode Discovery and Prior Fitting

Let $\mathcal{I}$ be the indices in the prior-fitting split with valid, unclipped brightness and shadow coordinates, matching Section 3.2. This same set is used for mode discovery and CVAE fitting; missing observations in the remaining coordinates are handled by Eq. (25). Learning all seven coordinates requires positive reconstruction support for each coordinate within $\mathcal{I}$. For $i \in \mathcal{I}$, the clustering features and weights are $\mathbf{u}_i = [\bar{b}_i, \bar{s}_i]^\top$ and $v_i = w_{i,b} w_{i,s}$, where $w_{i,b}$ and $w_{i,s}$ denote the brightness and shadow entries of the reconstruction weight vector. For each $K \in \{2, 3, 4, 5, 6\}$ with $K < |\mathcal{I}|$, weighted K-means minimizes

$$\sum_{i \in \mathcal{I}} v_i \min_{1 \le k \le K} \|\mathbf{u}_i - \mathbf{q}_k\|_2^2, \tag{27}$$

where $\mathbf{q}_k$ is a cluster center. We run 20 initializations per candidate and retain the lowest-objective solution with nonempty clusters. Let k_i be its assignment. For a sample whose cluster has another member, the weighted within-cluster and nearest-other-cluster distances are

$$\begin{aligned} A_i &= \frac{\sum_{i' \in \mathcal{I}:\, k_{i'} = k_i,\, i' \neq i} v_{i'} \|\mathbf{u}_i - \mathbf{u}_{i'}\|_2}{\sum_{i' \in \mathcal{I}:\, k_{i'} = k_i,\, i' \neq i} v_{i'}}, \\ B_i &= \min_{k \neq k_i} \frac{\sum_{i' \in \mathcal{I}:\, k_{i'} = k} v_{i'} \|\mathbf{u}_i - \mathbf{u}_{i'}\|_2}{\sum_{i' \in \mathcal{I}:\, k_{i'} = k} v_{i'}}. \end{aligned} \tag{28}$$

For nonsingleton clusters with $\max\{A_i, B_i\} > 0$, we compute a weighted variant of the silhouette score (Rousseeuw, 1987) and select K using

$$S_i = \frac{B_i - A_i}{\max\{A_i, B_i\}}, \qquad \mathrm{WSil}(K) = \frac{\sum_{i \in \mathcal{I}} v_i S_i}{\sum_{i \in \mathcal{I}} v_i}, \qquad K^\star = \arg\max_K \mathrm{WSil}(K), \tag{29}$$

with $S_i = 0$ for singleton clusters or $A_i = B_i = 0$, and ties broken in favor of fewer clusters. Candidates that cannot form K nonempty clusters are excluded. If no candidate is feasible, mode discovery cannot proceed. The selected modes describe brightness–shadow regimes; all seven coordinates are modeled jointly within each mode. For the selected assignments, the weighted mode frequencies used in Eq. (3) are

$$\pi_k = \frac{\sum_{i \in \mathcal{I}:\, k_i = k} v_i}{\sum_{i \in \mathcal{I}} v_i}, \qquad k = 1, \ldots, K^\star. \tag{30}$$

The CVAE encoder and decoder each have two hidden MLP layers with SiLU activations. The encoder takes the standardized coordinates, measurement confidences, and one-hot mode encoding specified in Section 3.2. Its posterior is $q_i = \mathcal{N}(\boldsymbol{\mu}_{\psi,i}, \mathrm{diag}(\exp \boldsymbol{\ell}_{\psi,i}))$, where $\boldsymbol{\mu}_{\psi,i}$ and $\boldsymbol{\ell}_{\psi,i}$ are the posterior mean and log variance. These latent-space quantities are distinct from the descriptor normalization statistics $\boldsymbol{\mu}, \boldsymbol{\sigma}$ in Eq. (26). Log variances are bounded to $[-12, 8]$, and posterior samples use $\mathbf{z}_i = \boldsymbol{\mu}_{\psi,i} + \exp(\boldsymbol{\ell}_{\psi,i}/2) \odot \boldsymbol{\zeta}_i$, with $\boldsymbol{\zeta}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For this diagonal Gaussian,

$$\frac{1}{d_z} D_{\mathrm{KL}}(q_i \| \mathcal{N}(\mathbf{0}, \mathbf{I})) = \frac{1}{2d_z} \sum_{h=1}^{d_z} [\exp(\ell_{\psi,ih}) + \mu_{\psi,ih}^2 - 1 - \ell_{\psi,ih}]. \tag{31}$$

The decoder reconstructs all seven standardized coordinates as $g_\theta(\mathbf{z}_i, \mathbf{e}_{k_i})$. The encoder and decoder are optimized over $\mathcal{I}$ using Eq. (2), with w_{ij} from Eq. (25) and the same $\beta, d_z, \epsilon_{\mathrm{rec}}$ as in the main text. The shared latent variable permits dependence among output coordinates; a factorized Gaussian latent prior does not require the seven decoded coordinates to be independent.

The learned prior is evaluated and selected on the grouped holdout using marginal coverage, brightness–shadow coverage, boundary-hit rates, and measurement validity. The holdout uses the fitting split's fixed normalization statistics and cluster centers, without refitting either. Missing or clipped reference coordinates are not treated as exact targets. No target validation or test images

enter descriptor normalization, mode discovery, CVAE fitting, or this holdout check. Priors are fitted separately for each target dataset. After selection, we freeze the fitted CVAE together with its normalization statistics $\boldsymbol{\mu}, \boldsymbol{\sigma}$, cluster centers $\mathbf{q}_k$, and mode probabilities $\boldsymbol{\pi}$. The fitting set $\mathcal{I}$ remains unchanged; the holdout is not added for a further fit. Thus Eq. (3) uses the frozen decoder and the normalization statistics from the same fitting split, as in Section 3.3.

A.3 Rendering Controls and Calibration

This appendix specifies the controls in Eq. (4). Scalar descriptor symbols in this section denote entries of the accepted sampled vector $\tilde{\boldsymbol{\phi}}$. We use the same brightness proxy, $\varepsilon = 1/255$, and $\delta = 10^{-6}$ as in Appendix A.1.1. Here H, W refer to the rendering resolution. The operator order is spatial modulation, channel gains, blur, exposure–shadow control, display-domain noise, and terminal quantization and JPEG encoding. All operators preserve pixel coordinates, so the source boxes and categories are retained.

A.3.1 Accepted Configurations and Rendering Targets

We sample modes directly as $k \sim \mathrm{Cat}(\boldsymbol{\pi})$, with π_k defined in Eq. (30), and sample and decode the latent variable as in Eq. (3). A candidate must satisfy the coordinate-wise bounds in Table 7 and the brightness–shadow compatibility constraints. In particular, the descriptor definitions imply the necessary quantile-feasibility conditions

$$0 < 2^b - \varepsilon \le 1, \qquad 0 \le 2^{b-s} - \varepsilon \le 2^b - \varepsilon. \tag{32}$$

These ensure a positive target median and an ordered, nonnegative lower quantile before numerical protection. Coordinate-wise bounds alone do not ensure this compatibility. They also do not guarantee that a particular source image can realize the requested statistics; realizability is checked after rendering. Invalid joint candidates are rejected and resampled as complete vectors, with at most 128 attempts. Exhausting this budget produces a sampling failure rather than clipping an invalid vector into the allowed range. The accepted parameter distribution is p_T, as in Section 3.3.

The protected brightness targets used by the renderer are

$$\begin{aligned} m_\star &= \mathrm{clip}_{[0.001,0.70]}(2^b - \varepsilon), \\ q_{10,\star} &= \max\{(m_\star + \varepsilon)2^{-s} - \varepsilon, 0\}. \end{aligned} \tag{33}$$

Here $m_\star$ is the target median and $q_{10,\star}$ is the corresponding lower-quantile target. These numerical guards do not replace rejection of incompatible samples. Since they can shift the exact inverse of the descriptors, terminal errors are checked against the original sampled b, s, not against newly defined descriptor values derived from the guarded targets.

A source-image identity, epoch, and base seed determine its sampled configuration. Once accepted, $\tilde{\boldsymbol{\phi}}$ stays fixed throughout exposure search and rendering retries. The random fields in $\boldsymbol{\xi}$ specify one realization; they are held fixed within a search and may change between rendering retries. Resampling a configuration across epochs is therefore separate from retrying its realization.

A.3.2 Spatial and Chromatic Controls

Given a smooth random field P with zero mean and unit standard deviation, the spatial operator $\mathcal{S}_a$ multiplies every linear-RGB channel by

$$F_{\mathrm{sp}}(p) = \frac{2^{aP(p)}}{(HW)^{-1}\sum_q 2^{aP(q)}}. \tag{34}$$

The field has mean gain one and becomes the identity when $a = 0$. This normalization does not preserve the mean or median brightness of every source image. Moreover, a measures coarse variation in display-domain brightness, whereas F_{sp} acts in linear RGB; the field amplitude controls spatial variation without guaranteeing that its terminal re-estimate equals a.

For $\mathcal{C}_{c_{RG},c_{BG}}$, the normalized channel-gain vector is

$$\mathbf{g} = \frac{(2^{c_{RG}}, 1, 2^{c_{BG}})^\top}{0.2126\,2^{c_{RG}} + 0.7152 + 0.0722\,2^{c_{BG}}}. \tag{35}$$

We apply g channel-wise in linear RGB. The normalization preserves the linear luminance of a neutral reference before clipping, not that of every colored pixel. The gains introduce the requested relative channel changes; pre-existing source chromatic bias and subsequent processing can affect the measured output bias. Exposure–shadow control follows both spatial and chromatic changes so that brightness is adjusted after these operations.

A.3.3 Blur and Noise Calibration

We map r and n to rendering strengths by monotone piecewise-linear interpolation of the fixed knots in Table 8. The reference Gaussian width $\sigma_{\mathrm{ref}}(r)$ is expressed in pixels at the analysis scale of Appendix A.1.1. Since that preprocessing only downsamples images larger than 512 pixels on the longer side, the rendering width is

$$\sigma_{\mathrm{blur}} = \sigma_{\mathrm{ref}}(r) \max\left\{1, \frac{\max(H, W)}{512}\right\}. \tag{36}$$

Gaussian filtering uses reflection padding; a zero width gives the identity operator. The width is a rendering control rather than the descriptor r itself. The same distinction applies to the noise amplitude $\rho(n)$ and the residual statistic n. These mappings guide appearance synthesis but are not exact inverses of the estimators for every source scene.

Table 8: Fixed control knots for the observable blur and noise descriptors. Gaussian widths refer to the analysis scale.

r	0	.0262	.0593	.1418	.2001	.4284	.6195	.8585	.9206	.9751	1
σ_{ref}	0	0	.5	.75	1	1.5	2	3	4	5	6

n	0	.0023	.0083	.0243	.0368	.0496	.08
$\rho(n)$	0	.010	.020	.040	.060	.080	.162

Noise is added to the clipped display-RGB output x_{tone} of the exposure–shadow operator $\mathcal{T}_{b,s}$. Its brightness proxy Y_{tone} is computed after this channel clipping and before noise addition, as defined in Eq. (40). For channel $\chi \in \{R, G, B\}$ and pixel p, we use

$$\begin{aligned}\eta_\chi(p) = \rho(n)\big[&\sqrt{0.40^2 + 0.16^2}\,\eta_{\mathrm{ind},\chi}(p) \\ &+ 0.52\sqrt{\max\{Y_{\mathrm{tone}}(p), \varepsilon\}}\,\eta_{\mathrm{shot},\chi}(p) + 0.08\eta_{\mathrm{row},\chi}(p_y)\big].\end{aligned} \tag{37}$$

The three fields are independent standard Gaussians, sampled independently across channels. The first two are independent across pixels; the row component is sampled per row and shared across its columns. The signal-independent coefficient combines two independent Gaussian contributions in variance, without treating them as separately identifiable noise mechanisms. The luminance-dependent and signal-independent terms give a Poisson–Gaussian-inspired appearance model in display RGB (Foi et al., 2008), not a calibrated sensor-noise model. The noise operator $\mathcal{N}_n$ outputs $\mathrm{clip}_{[0,1]}((x_{\mathrm{tone}})_\chi(p) + \eta_\chi(p))$ in each channel. We then quantize to 8 bits and JPEG-encode at quality 80 with chroma subsampling setting 2, implementing $\mathcal{J}$ in Eq. (4).

A.3.4 Exposure–Shadow Mapping

Let x_{lin} denote linear RGB after spatial, chromatic, and blur processing, and let Γ be the channel-wise sRGB encoding function inverse to Eq. (18). An exposure control e produces

$$x_e = \Gamma\Big(\mathrm{clip}_{[0,1]}(2^e x_{\mathrm{lin}})\Big), \qquad Y_e = 0.2126(x_e)_R + 0.7152(x_e)_G + 0.0722(x_e)_B. \tag{38}$$

Clipping and sRGB encoding are applied to RGB channels before forming Y_e; encoding a scalar linear-luminance average is not equivalent for general colored pixels. The internal exposure control e is distinct from the sampled display-brightness coordinate b.

Let $m = Q_{0.50}(Y_e)$ and $s_e = \log_2[(m+\varepsilon)/(Q_{0.10}(Y_e)+\varepsilon)]$. For a positive median, the shadow mapping is

$$M_{\gamma,m}(Y_e) = \mathrm{clip}_{[0,1]}\left[(m+\varepsilon)\left(\frac{Y_e+\varepsilon}{m+\varepsilon}\right)^{\gamma} - \varepsilon\right], \qquad \gamma = \mathrm{clip}_{[0.08,10]}\left(\frac{s}{\max\{s_e,\delta\}}\right). \tag{39}$$

The scalar mapping satisfies $M_{\gamma,m}(m) = m$. When clipping is inactive, it scales the log-quantile gap approximately by γ; bounding γ can prevent exact attainment of s. The tone operator $\mathcal{T}_{b,s}$ scales the display-RGB channels together and clips them before noise addition:

$$x_{\mathrm{tone}}(p) = \begin{cases} \mathrm{clip}_{[0,1]}\left[x_e(p)\dfrac{M_{\gamma,m}(Y_e(p))}{Y_e(p)}\right], & Y_e(p) > 0, \\ \mathbf{0}, & Y_e(p) = 0, \end{cases} \tag{40}$$
$$Y_{\mathrm{tone}}(p) = 0.2126(x_{\mathrm{tone}})_R(p) + 0.7152(x_{\mathrm{tone}})_G(p) + 0.0722(x_{\mathrm{tone}})_B(p).$$

Clipping acts channel-wise, and zero-luma pixels remain zero. Before this channel clipping, the common RGB scaling realizes the scalar brightness mapping at positive-luma pixels without an additive denominator offset that would darken the median pivot. The noise operator uses the post-clipping Y_{tone} in Eq. (37), adds noise to x_{tone}, and clips again before terminal quantization and encoding. Channel clipping, zero-luma pixels, noise, and JPEG encoding can still change the terminal statistics, so the complete renderer requires the checks below.

Exposure is searched over $[-16, 4]$ to approach $m_\star$ in Eq. (33), initially at a proxy resolution with a longer side of at most 192 pixels, and followed by at most two finite-difference corrections. Bisection requires a valid bracket for the terminal-median residual with $\tilde{\phi}$ and $\boldsymbol{\xi}$ fixed. The pivot m and current gap s_e are recomputed for each exposure candidate. The search targets the median after the remaining processing steps, rather than assuming that noise and encoding preserve it. Undefined statistics or a failed bracket invalidate the attempt, and the final output is accepted only after full-resolution rendering and terminal evaluation at the common analysis scale.

Figure 9 illustrates the joint effect of the brightness and shadow controls on the same scene. Increasing b raises overall brightness, while increasing s changes the separation between mid-tones and shadows through Eq. (39).

A.3.5 Terminal Checks and Failure Handling

We decode the terminal JPEG and measure its brightness proxy Y_{out} using the same analysis preprocessing as the target descriptors. The nominal acceptance checks are

$$|\log_2(Q_{0.50}(Y_{\mathrm{out}})+\varepsilon) - b| \le 0.45, \qquad \left|\log_2\frac{Q_{0.50}(Y_{\mathrm{out}})+\varepsilon}{Q_{0.10}(Y_{\mathrm{out}})+\varepsilon} - s\right| \le 0.75. \tag{41}$$

The estimates must satisfy the brightness and shadow validity rules in Appendix A.1.2. Errors are computed from raw estimates before descriptor bounding, so clipping cannot conceal a mismatch. These tolerances are in the defined display-domain log coordinates, not physical exposure units. If relaxed tolerances are used, their values must be recorded and may not exceed 1.25 for brightness or 1.05 for shadow contrast. Outputs must retain the source geometry and have valid dimensions and encoding; neither the fraction of fully black pixels nor that of fully white pixels may reach 0.9995.

Any bounded terminal shadow correction is followed by renewed encoding and both descriptor checks, since a shadow adjustment can also change brightness. At most four stochastic realizations are attempted for a fixed $\tilde{\phi}$. Exhausting these retries produces an explicit rendering failure, separate from the 128-attempt parameter-sampling budget. Failed outputs are not admitted to $\mathcal{D}_{\mathrm{syn}}$. Acceptance can depend on the source scene and the requested degradation, so the distribution of configurations represented among accepted training pairs can differ from p_T. Only brightness and shadow contrast are explicitly checked against sampled targets; the remaining coordinates guide synthesis through the controls above.



Figure 9: Rendering examples for joint brightness and shadow control. Rows vary the requested brightness coordinate $b \in \{-5, -4, -3\}$, and columns vary the requested shadow-contrast coordinate $s \in \{0.5, 1.5, 2.5\}$. The labels B and S in the panels correspond to b and s, respectively; the EV notation denotes increments in the defined base-2 display-luma coordinates rather than calibrated physical exposure. The scene geometry is preserved across the different appearance settings.

A.4 Ablation of Degradation Dimensions

Table 9 evaluates the contribution of each degradation dimension by removing it individually on ExDark. The full model achieves 80.9% mAP50, and every removal reduces performance. Removing brightness causes the largest drop of 2.9 percentage points, followed by spatial illumination variation at 2.4 points. Removing visible noise or either chromatic bias reduces mAP50 by 1.9 points, while removing blur and shadow contrast yields smaller drops of 1.1 and 0.9 points, respectively. These results support the complementary contributions of all seven dimensions, with brightness and spatial illumination variation having the largest effects in this comparison.

Table 9: Single-dimension ablations on ExDark. Δ denotes the mAP50 of each ablated variant minus that of the full model, in percentage points (pp); negative values indicate performance degradation.

Degradation configuration	**mAP50 (%)**	**Δ (pp)**
CoDeS (all seven dimensions)	**80.9**	0.0
w/o brightness (b)	78.0	-2.9
w/o shadow contrast (s)	80.0	-0.9
w/o visible noise (n)	79.0	-1.9
w/o blur (r)	79.8	-1.1
w/o red–green bias (c_{RG})	79.0	-1.9
w/o blue–green bias (c_{BG})	79.0	-1.9
w/o spatial illumination variation (a)	78.5	-2.4

A.5 Effectiveness of Synthesis Across Detectors

To examine whether the utility of our synthesized low-light images extends beyond SAM 3, we additionally train YOLOE-11 and RF-DETR with images generated by our synthesis method and evaluate on ExDark. Table 10 compares each detector's baseline without our synthetic training data with its result after training with the synthesized images. For SAM 3, we reproduce the baseline and *CVAE Generation* results without the LFA from Table 5, so that this comparison does not include the additional benefit of feature adaptation.

Table 10: ExDark results (mAP50, %).

Detector	Baseline	With synthesis	Gain
YOLOE-11	62.3	**65.6**	+3.3
RF-DETR	72.5	**73.6**	+1.1
SAM 3	77.7	**79.6**	+1.9

Training with our synthesized images increases mAP50 from 62.3% to 65.6% for YOLOE-11 and from 72.5% to 73.6% for RF-DETR, corresponding to gains of 3.3 and 1.1 percentage points, respectively. SAM 3 improves from 77.7% to 79.6% without the LFA, a gain of 1.9 points. The positive within-detector gains across all three models support the usefulness of the synthesized images for low-light detection beyond our SAM 3-based implementation. Since synthesis operates on training images and preserves their annotations, it can provide supervision for different detectors without requiring a detector-specific synthesis module.

A.6 Experimental Configuration

Degradation prior. We fit each domain's CVAE with batch size 128 and learning rate 10^{-3}, using the settings in Table 11. A grouped 20% internal holdout supports prior evaluation and selection. Descriptor normalization and illumination-mode discovery use only the remaining prior-fitting split, and CVAE training uses its eligible indices $\mathcal{I}$, as specified in Appendix A.2. The selected CVAE and its associated normalization statistics, cluster centers, and mode probabilities are then frozen for synthesis; the holdout is not included in a subsequent fit. Each source image contributes one training view per epoch: normal-light pixels are retained with probability 0.1, and a low-light view is synthesized otherwise; both use the source annotations.

Table 11: Dataset-specific prior and detector settings. The LFA learning rate is 10^{-4} for all three datasets.

Dataset	Latent dim.	Hidden dim.	β	Prior epochs	Detector LR
ExDark	4	128	0.08	200	1.25×10^{-4}
NOD	8	192	0.04	250	1.25×10^{-4}
DARK FACE	8	128	0.08	250	6.25×10^{-6}

Detector optimization. The detection module and LFA use AdamW, batch size 3 without gradient accumulation, and weight decay 0.05, excluding bias, normalization, and gating parameters. The inverse-square-root schedule uses timescale/warmup/cooldown values of 400/800/800. Runs use seed 20260621 on one RTX 3090 for up to eight epochs. Training samples square resize sizes from $480, 512, \ldots, 992$ and pads to 1008×1008; validation and test use a fixed 1008×1008 resize. RGB normalization uses mean and standard deviation 0.5.

Checkpoint selection. We select the checkpoint with the highest validation mAP50; on ExDark, the best performance is typically reached after the first epoch.