DARKOCC: ILLUMINATION-ROBUST INDOOR OCCUPANCY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing RGB-based occupancy models typically assume well-illuminated inputs and suffer substantial performance degradation under low-light conditions. Image enhancement only partially addresses this issue, while jointly training on normal- and low-light images can compromise performance under normal illumination. We present **DarkOcc**, a framework for robust indoor occupancy prediction across illumination conditions. DarkOcc first learns illumination-robust visual features through multi-level feature adaptation to improve occupancy prediction under low-light conditions. A geometry-guided occupancy completion module then combines these features with base depth and voxel predictions to recover missing occupied structure, addressing occupancy incompleteness exacerbated by low-light degradation. On Occ-ScanNet-mini, DarkOcc improves Geo-IoU over GPOcc trained on mixed-light data (MLA) by **5.57 pp** under severe low light and **5.13 pp** under normal illumination. Further experiments demonstrate its effectiveness in embodied occupancy prediction. We also construct a real-world indoor occupancy benchmark covering normal and low-light conditions. Under low illumination, DarkOcc improves Geo-IoU over MLA by **9.76 pp**.

Keywords: Indoor 3D Occupancy Prediction; Low-Light Robustness; Feature Adaptation; Geometry-Guided Occupancy Completion; Synthetic-to-Real Generalization

1 INTRODUCTION

RGB-based indoor occupancy prediction infers the 3D distribution of occupied and free space, providing essential spatial information for applications such as robot navigation and manipulation (Wu et al., 2025b; Jiang et al., 2026; Wang et al., 2026). Existing methods, however, typically assume well-illuminated inputs. In general environments, robots may encounter low-light conditions where object boundaries, textures, and local structural details become difficult to perceive (Zeng et al., 2025; Yao & Chuah, 2025). Such visual degradation can lead to inaccurate or inconsistent occupancy estimates.

A simple way to mitigate this problem is to enhance low-light images before feeding them into the occupancy prediction model, using image enhancement methods such as LCDPNet (Wang et al., 2022), CSEC (Li et al., 2024), HVI-CIDNet (Yan et al., 2025), and Multinex (Brateanu et al., 2026). However, this strategy only partially recovers occupancy accuracy, as the enhancement process is optimized independently of the downstream occupancy objective, and therefore may fail to recover the visual cues that are most relevant for occupancy reasoning. We also find that jointly training on normal- and low-light images can substantially improve low-light performance. However, this improvement comes at the cost of accuracy under normal illumination. This motivates our central question: *How can a single RGB-based occupancy model improve robustness to low illumination without sacrificing accuracy under normal illumination?*

We make two important observations from controlled experiments. First, beyond simple mixed training under varying illumination conditions, more targeted feature adaptation strategies can learn more robust visual representations and improve occupancy prediction. However, improving visual representations alone does not fully resolve the performance degradation issue caused by insufficient illumination. Even after feature adaptation, stronger low-light degradation causes more occupied voxels that were correctly predicted under normal light to be misclassified as free space (Fig. 1).

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.

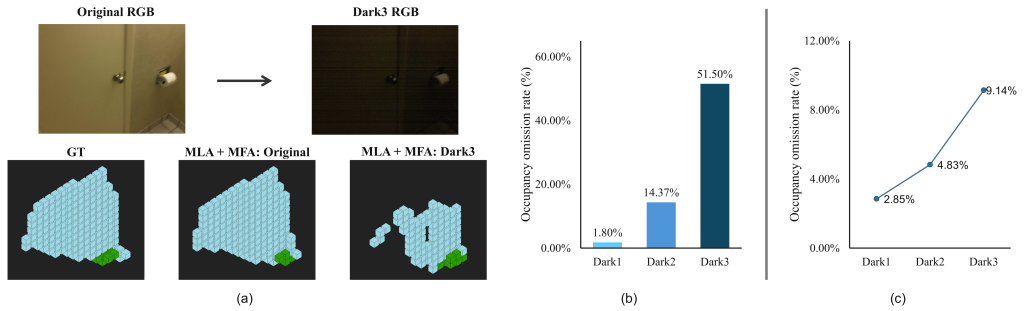


Figure 1: Low light exacerbates occupancy incompleteness after feature adaptation. All results use mixed-light augmentation (MLA) and multi-level feature adaptation (MFA), without GOC. (a) Predictions for the same scene under Original and Dark3. (b) For the frame in (a), each Dark1–Dark3 rate is the number of ground-truth occupied voxels predicted as free divided by the total number of ground-truth occupied voxels in the evaluation domain. (c) Dataset-level statistics on Occ-ScanNet-mini, measuring occupied voxels correctly recovered under Original but misclassified as free under low light.

Across the Occ-ScanNet-mini test set, the proportion of these voxels increases from 2.85% under mild low-light degradation (Dark1) to 9.14% under severe degradation (Dark3). This occurs because RGB-based occupancy prediction requires not only reliable visible-surface features, but also effective volumetric reasoning beyond observed surfaces to recover complete 3D structures (Zhou et al., 2026a;c; Sun et al., 2025).

Building on these observations, we present DarkOcc, an indoor occupancy prediction framework designed for varying illumination conditions. DarkOcc first learns illumination-robust, geometry-aware representations through constrained multi-level feature adaptation. To recover occupied structure still missing after feature adaptation, we further introduce a geometry-conditioned occupancy repair network. Specifically, Geometry-Guided Occupancy Completion (GOC) uses base depth to associate image features with voxel locations and combines them with base occupancy predictions. Local and coarse 3D context then guide residual corrections to recover missing occupied structure.

We evaluate DarkOcc under four illumination conditions: the unmodified normal-light input (Original) and three synthetic low-light levels of increasing severity (Dark1–Dark3). Given the cost of acquiring real-world 3D occupancy annotations, we generate all low-light training inputs by synthetically degrading normal-light training images. For real-world evaluation, we create an indoor low-light dataset with OccScanNet-compatible 3D occupancy ground truth and establish a unified evaluation benchmark. The real low-light dataset is used exclusively for testing to assess synthetic-to-real transfer.

Across the mini and full Occ-ScanNet settings and the EmbodiedOcc-ScanNet-mini benchmark, DarkOcc improves performance under both normal and low-light conditions. On Occ-ScanNet-mini, DarkOcc improves Geo-IoU from **35.650%** for the original GPOcc to **61.958% (+26.31 pp)** under the most severe synthetic low-light setting. Compared with GPOcc trained on mixed-light data (MLA), the gains are **5.57 pp** under severe low light and **5.13 pp** under normal illumination. To complement the synthetic experiments, we collect a real-world indoor occupancy dataset under both normal and low illumination. On the real low-light data, DarkOcc outperforms MLA by **9.76 pp** in Geo-IoU.

Our main contributions are as follows:

1. **A systematic study of low-light indoor occupancy prediction.** We systematically study indoor RGB occupancy prediction under low-light degradation. Controlled experiments reveal that low light amplifies occupancy omissions that persist after feature adaptation, motivating explicit occupancy repair.
2. **A unified framework for illumination-robust occupancy prediction.** We propose DarkOcc, which combines multi-level feature adaptation with Geometry-Guided Occupancy Completion. The framework adapts visual representations and selectively recovers missing occupancy while

preserving existing occupied predictions, improving accuracy under both normal and low illumination.

3. **A real low-light dataset and benchmark.** We construct an indoor low-light dataset with OccScanNet-compatible 3D occupancy ground truth. Evaluations spanning controlled synthetic degradation, real low-light tests, and embodied experiments assess robustness across illumination conditions and synthetic-to-real transfer.

2 Related Work

Indoor occupancy prediction. Indoor occupancy prediction recovers volumetric geometry and semantics from partial visual observations (Chen et al., 2025; Zhou et al., 2026b). ISO (Yu et al., 2024) uses depth-guided voxel lifting, while SliceOcc (Li et al., 2025a) adopts vertical slice representations. EmbodiedOcc (Wu et al., 2025b) extends prediction to embodied scene understanding. Gaussian-based methods improve efficiency through depth-guided initialization in SplatSSC (Qian et al., 2026), visual geometry priors in GPOcc (Zhou et al., 2026a), and likelihood-based primitive learning in FLM-Occ (Chen et al., 2026). GPOcc++ (Zhou et al., 2026c) unifies multiple views and temporal observations. These advances primarily concern scene representation and geometric reasoning. We study their vulnerability to low-light degradation and address occupancy incompleteness that remains after feature adaptation.

Low-light image enhancement. Low-light enhancement restores visibility while addressing noise, color distortion, and exposure variation (Xu et al., 2025; Liu et al., 2026). Representative approaches use signal-quality guidance (Xu et al., 2022), illumination decomposition (Cai et al., 2023), and color-aware processing (Li et al., 2024; Yan et al., 2025); DarkIR (Feijoo et al., 2025) jointly handles low illumination, noise, and blur. Such methods offer useful preprocessing for visual perception, but image restoration objectives do not directly supervise 3D occupancy. We instead target occupancy accuracy across normal and low illumination within the prediction model.

Low-light 3D perception and reconstruction. RawNeRF (Mildenhall et al., 2022) reconstructs scenes from noisy raw images, while DarkGS (Zhang et al., 2024), RoSe (Li et al., 2025c), and SplatBright (Wen et al., 2025) address reconstruction and illumination modeling with Gaussian representations. Beyond reconstruction, DepthDark (Zeng et al., 2025) studies monocular depth estimation in low light. LIAR (Wu et al., 2025a) uses illumination cues to guide feature sampling and BEV refinement. Building on GPOcc's Gaussian occupancy pipeline (Zhou et al., 2026a), we adapt multiple encoder levels through illumination-conditioned residuals whose norms are bounded relative to the input features. Occupied structure still missing after adaptation motivates geometry-guided completion restricted to voxels predicted as free.

3 Method

3.1 Problem Formulation and Failure Modes

We study RGB-based indoor occupancy prediction across normal and low illumination. Given an image I and camera parameters, the model predicts a class-score vector z_v^{base} at each voxel v in a local volume $\mathcal{V}$, distinguishing occupied semantic categories from free space. Embodied occupancy prediction accumulates successive observations into a global scene representation (Wu et al., 2025b; Guo et al., 2026; Wang et al., 2026).

Low-light degradation affects this process at two levels. First, reduced brightness and contrast obscure boundaries and textures, weakening the backbone's structural and semantic representations before occupancy construction (Zeng et al., 2025; Ou et al., 2026). Second, insufficient evidence can cause occupied regions to be assigned to free space. For a known ground-truth label y_v, an occupied voxel is misclassified as free space when

$$y_v \in \mathcal{C}_{\text{occ}}, \qquad \arg\max_c z_{v,c}^{\text{base}} = \text{free}, \tag{1}$$

where $\mathcal{C}_{\text{occ}}$ contains the occupied semantic classes. This definition describes the failure mode for analysis; ground-truth labels are not used to select corrections at inference.

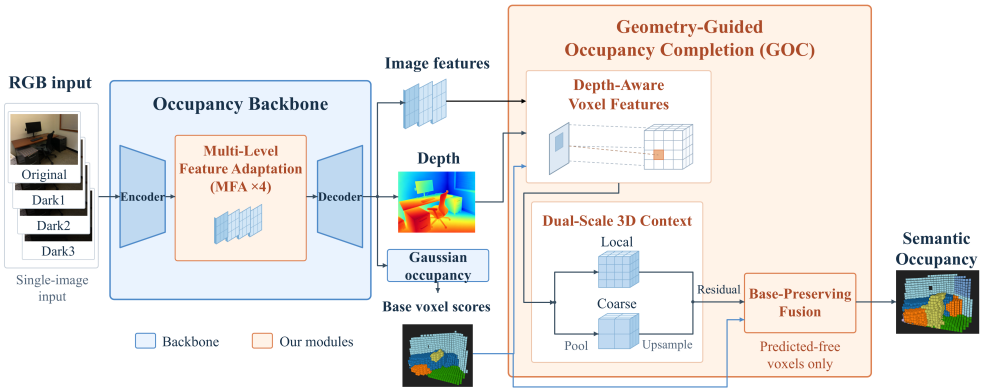


Figure 2: Overview of DarkOcc for indoor semantic occupancy prediction. Each prediction uses a single RGB image under normal illumination (Original) or one of three increasingly severe synthetic low-light conditions (Dark1–Dark3). Multi-Level Feature Adaptation (MFA) adjusts four encoder feature levels using brightness and contrast cues. The backbone produces image features, predicted depth, and Gaussian occupancy, from which base voxel scores are obtained. Geometry-Guided Occupancy Completion (GOC) combines image features, predicted depth, and base voxel scores to construct depth-aware voxel features. Local and pooled coarse-scale 3D branches then predict residual corrections, which are added to the base scores only at voxels initially predicted as free space to obtain the final semantic occupancy. Blue and orange indicate the backbone and our proposed modules, respectively.

Occupancy incompleteness occurs under normal light but is exacerbated by low light (Fig. 1). DarkOcc addresses both stages: MFA adapts visual representations before occupancy construction, while GOC recovers missing occupied structure afterward.

3.2 DarkOcc Framework

Figure 2 presents DarkOcc, built on GPOcc (Zhou et al., 2026a) with a VGGT encoder (Wang et al., 2025b) and DPT decoder (Ranftl et al., 2021). Multi-Level Feature Adaptation (MFA) adapts four encoder interfaces. The backbone produces decoded image features, depth, and Gaussian occupancy, which yields base voxel scores z_v^{base}.

Geometry-Guided Occupancy Completion (GOC) combines decoded image features, base depth, and voxel scores with local and coarse 3D context to complete missing occupied structure. Inference uses only RGB and camera parameters, with the embodied extension described in Section 3.5.

3.3 Multi-Level Feature Adaptation

3.3.1 Multi-level residual adaptation

MFA places residual adapters at the four encoder interfaces used by DPT (Ranftl et al., 2021; Wang et al., 2025b) to address corruption across feature levels. At interface ℓ, F_ℓ and $\widetilde{F}_\ell$ denote the original and adapted spatial patch features. Each adapter applies normalization, channel compression, depthwise spatial convolution, and projection back to the interface dimension. Camera and register tokens remain unchanged.

3.3.2 Illumination-conditioned modulation

Local luminance, dark-pixel fraction, and local contrast condition spatial adapter responses, while global mean brightness controls their overall strength. These gates modulate residuals before feature addition, making adaptation depend on image degradation without introducing a separate prediction objective.

3.3.3 BOUNDED ADAPTATION

We scale the modulated adapter response U_ℓ to bound its contribution relative to the current encoder feature F_ℓ:

$$\widetilde{F}_\ell = F_\ell + \eta_\ell U_\ell, \qquad \eta_\ell = \min\left(1, \frac{\rho_F \|F_\ell\|_F}{\max(\|U_\ell\|_F, \epsilon)}\right). \tag{2}$$

Here ρ_F bounds the residual relative to the input feature norm, and ϵ ensures numerical stability. Per-image norms span spatial positions and channels, ensuring $\|\widetilde{F}_\ell - F_\ell\|_F \le \rho_F \|F_\ell\|_F$. This bound applies to the adapter contribution in each forward pass, not to changes in F_ℓ caused by encoder optimization. Zero initialization starts each adapter as an identity mapping. Appendix A gives gate definitions and numerical settings.

3.4 GEOMETRY-GUIDED OCCUPANCY COMPLETION

GOC operates under all illumination conditions. Its editable region depends on base occupancy predictions and valid image projections, rather than an image-level dark/light decision.

3.4.1 DEPTH-AWARE VOXEL FEATURE CONSTRUCTION

Appearance alone cannot distinguish voxels along a viewing ray, so we use base depth as a geometric reference. A two-dimensional ray grid samples the image plane, with camera intrinsics defining each ray. An image adapter maps decoded features to descriptors f_i, and a width predictor outputs positive influence scales τ_i at each grid location i. Projecting voxel v into the grid samples $\bar{f}_v$, $\bar{\tau}_v$, and the base camera-depth estimate $\bar{Z}_v^{\text{base}}$. The appearance cue becomes

$$\widetilde{f}_v = (1 + n_v)\bar{f}_v, \qquad n_v = \exp\left[-\frac{(Z_v - \bar{Z}_v^{\text{base}})^2}{2\bar{\tau}_v^2}\right]. \tag{3}$$

Here Z_v is the voxel's camera depth and bars denote image-plane sampling. The proximity term strengthens appearance evidence near the predicted surface, while the residual factor retains evidence farther away. The learned width controls this modulation without replacing base depth or changing Gaussian scale.

3.4.2 DUAL-SCALE 3D CONTEXTUAL REASONING

For each voxel, X_v concatenates $\widetilde{f}_v$, base class scores z_v^{base}, and depth/projection cues. A pointwise projection maps X_v to a compact representation. The local 3D branch captures nearby boundaries and geometric continuity; a pooled coarse branch captures broader context and upsamples its output to voxel resolution. Depthwise 3D residual blocks keep both branches compact.

The branches predict a score residual r_v over occupied categories and free space. Keeping free space as an output allows GOC to distinguish missing structure from genuinely empty voxels.

3.4.3 BASE-PRESERVING RESIDUAL FUSION

We restrict editing to the set $\mathcal{E}$ of voxels predicted free by the base with valid image projections. The final scores are

$$z_v^{\text{out}} = \begin{cases} z_v^{\text{base}} + r_v, & v \in \mathcal{E}, \\ z_v^{\text{base}}, & v \notin \mathcal{E}. \end{cases} \tag{4}$$

Existing occupied predictions and their errors remain unchanged; eligible free voxels may remain free or become occupied. This constraint applies at inference; training optimizes the complete model. Appendix A details supervision and implementation.

3.5 EMBODIED OCCUPANCY PREDICTION

For embodied occupancy prediction, GPOcc's incremental Gaussian memory (Zhou et al., 2026a) stores framewise Gaussians, while a separate voxel field accumulates signed differences between completed and base voxel scores. Residuals are mapped to the global grid and averaged over valid observations, including zero-residual views. The mean is added only where the global base predicts

free space, without feedback into Gaussian memory. Mapping, aggregation, and sequence handling are detailed in Appendix B.1.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate monocular and embodied occupancy prediction under controlled low-light degradation and captured indoor illumination.

Datasets. We evaluate monocular prediction on Occ-ScanNet (Yu et al., 2024), derived from ScanNet (Dai et al., 2017). Each frame has a $60 \times 60 \times 36$ voxel volume spanning $4.8 \times 4.8 \times 2.88$ m, with 11 occupied categories and free space. Full Occ-ScanNet uses 45,610 training and 19,705 test images; Occ-ScanNet-mini uses 5,504 and 2,376, respectively. Each test set is evaluated under Original and Dark1–Dark3. EmbodiedOcc-ScanNet-mini (Wu et al., 2025b) contains 16 sequences of 30 posed frames, evaluated over the accumulated explored region at 8 cm resolution. Appendix B details the evaluation domains.

The real-data evaluation uses 8,208 frames from 19 recordings in 10 scenes: 3,222 under normal light and 4,986 under low light. Occupancy references are constructed through offline camera tracking, multi-frame depth fusion, semantic association, and reviewed geometric completion. Unsupported regions retain unknown labels, while ray tracing identifies observed free space. Only RGB and camera parameters are supplied to the models. Data statistics, annotation procedures, and the evaluation protocol are given in Appendix E.

Illumination conditions and metrics. Original uses unmodified RGB; Dark1–Dark3 apply increasing exposure reduction and noise to the same frames without changing geometry or labels. All methods receive identical cached inputs. We report Geo-IoU for occupied space and semantic mIoU over 11 occupied categories from pooled confusion counts (Wu et al., 2025b; Zhou et al., 2026a). Cells give Geo-IoU / mIoU in percent; dashes indicate unavailable measurements. Synthetic comparisons share target labels, evaluation masks, and metric computation (Appendix B); real-data settings appear in Appendix E. Appendices C and D describe synthesis and unseen degradations.

Implementation details. Inputs are resized to 392×518, with 16 Gaussian samples per ray. We train each complete configuration for 30 epochs on its corresponding Occ-ScanNet training split using AdamW (Loshchilov & Hutter, 2019), weight decay 0.01, linear warmup, cosine learning-rate decay, and gradient clipping. Experiments use four NVIDIA RTX 8000 GPUs and four NVIDIA L40S GPUs. Architecture, optimization, and timing details appear in Appendices A and F.

Compared methods and training protocol. Our baselines include EmbodiedOcc (Wu et al., 2025b), EmbodiedOcc++ (Wang et al., 2025a), SplatSSC (Qian et al., 2026), DiScene (Li et al., 2025b), GPOcc (Zhou et al., 2026a), and FLM-Occ (Chen et al., 2026) (64 Gaussians). For methods using an RGB-derived depth prior, the prior is recomputed from the input under each illumination condition. We evaluate these baselines using their original checkpoints. Mixed-light augmentation (MLA) trains GPOcc with Original and synthetic low-light images. Adding MFA yields the illumination-adapted base (IAB), and adding GOC yields DarkOcc. All three configurations use their final 30-epoch checkpoints for synthetic and real-data evaluation, without additional training on real data. The monocular and embodied comparisons in Tables 1 and 2 therefore evaluate complete systems under a common evaluation protocol with different training regimes; comparisons against MLA assess gains beyond mixed-light training.

4.2 MONOCULAR OCCUPANCY PREDICTION

Results on Occ-ScanNet-mini. Table 1 shows that DarkOcc outperforms the evaluated baselines across illumination levels. On Dark3, it exceeds the original GPOcc by 26.31 Geo-IoU / 25.15 mIoU points, combining training and method changes. Relative to MLA, the gains are 5.13 / 2.80 points on Original and 5.57 / 2.91 on Dark3. MLA already reduces the Original-to-Dark3 Geo-IoU gap from 27.51 to 2.68 points; DarkOcc further narrows it to 2.24 points. Thus, the main benefit beyond

Table 1: Monocular prediction performance on Occ-ScanNet-mini and Occ-ScanNet.

Method	Year–Venue	Occ-ScanNet-mini				Occ-ScanNet			
		Original	Dark1	Dark2	Dark3	Original	Dark1	Dark2	Dark3
EmbodiedOcc	2025-ICCV	52.03 / 44.39	47.44 / 39.72	38.75 / 30.41	24.54 / 16.32	51.58 / 42.67	46.76 / 37.82	38.07 / 29.34	24.27 / 17.03
EmbodiedOcc++	2025-ACM MM	55.88 / 48.20	51.32 / 43.60	41.95 / 34.07	26.32 / 18.96	-	-	-	-
DiScene	2025-RA-L	51.85 / 45.75	47.53 / 39.68	38.91 / 27.99	24.62 / 12.09	52.02 / 47.19	47.34 / 40.03	38.34 / 28.14	24.20 / 12.39
SplatSSC	2026-AAAI	60.85 / 48.48	54.35 / 43.35	43.07 / 32.68	25.89 / 17.01	62.84 / 51.84	56.09 / 45.63	43.74 / 33.80	25.69 / 17.96
FLM-Occ	2026-ECCV	61.05 / 53.10	59.64 / 51.81	53.28 / 44.90	36.86 / 27.23	64.78 / 56.57	62.32 / 53.91	53.43 / 44.99	36.54 / 27.02
GPOcc	2026-CVPR	63.16 / 55.85	60.71 / 53.63	52.84 / 45.61	35.65 / 28.58	63.54 / 56.34	60.92 / 53.14	53.32 / 44.95	38.77 / 29.86
MLA	-	59.07 / 54.55	58.76 / 54.01	58.46 / 53.55	56.39 / 50.82	61.15 / 55.48	59.87 / 53.21	58.93 / 52.45	54.17 / 45.49
DarkOcc	-	**64.20 / 57.35**	**64.10 / 57.26**	**63.67 / 56.48**	**61.96 / 53.73**	**66.38 / 58.71**	**65.88 / 58.07**	**64.28 / 56.01**	**59.29 / 49.58**

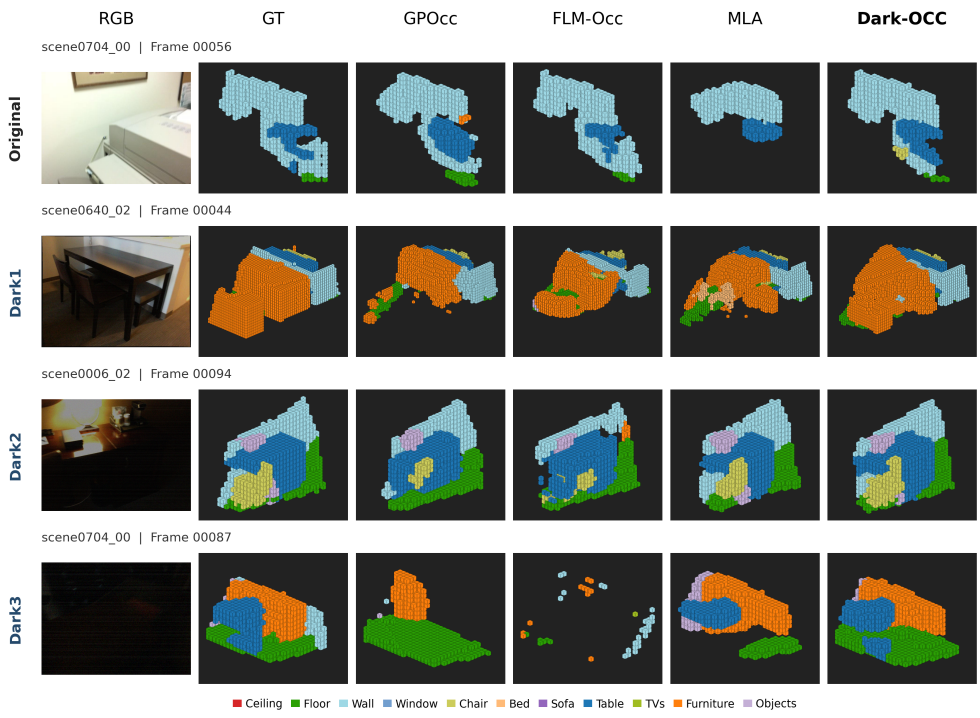


Figure 3: Qualitative comparison of indoor occupancy prediction under Original and Dark1–Dark3 conditions. Each row shows a different input frame, with the RGB input, ground-truth occupancy, and predictions from GPOcc, FLM-Occ, MLA, and DarkOcc. Colors denote semantic categories.

mixed-light training is higher absolute accuracy while retaining a small degradation gap. Section 4.5 separates the contributions of MFA and GOC.

Results on Occ-ScanNet. Table 1 extends the comparison to larger training and test sets. DarkOcc achieves 66.38 / 58.71 on Original and 59.29 / 49.58 on Dark3, outperforming the evaluated baselines across all four conditions. On Original, it exceeds FLM-Occ by 1.60 / 2.14 points. Relative to MLA, the gains are 5.23 / 3.23 points on Original and 5.12 / 4.09 points on Dark3. The Original-to-Dark3 Geo-IoU gap remains comparable to MLA, at 7.09 versus 6.98 points, so the principal benefit at this scale is higher absolute accuracy under both conditions. Class-wise results are reported in Appendix G.

Qualitative results. In Fig. 3, GPOcc and FLM-Occ miss substantial structure under Dark3. Mixed-light training recovers much of the scene, but visible gaps remain. DarkOcc produces more complete occupancy consistent with the ground truth and retains coherent structure under normal light, complementing the quantitative gains over MLA.

4.3 EMBODIED OCCUPANCY PREDICTION

All methods process the same 30 posed observations per EmbodiedOcc-ScanNet-mini sequence and are evaluated on the final accumulated global target. DarkOcc uses the fusion in Section 3.5; evaluation and aggregation settings are given in Appendix B.1.

Table 2: Embodied prediction performance on EmbodiedOcc-ScanNet-mini.

Method	Year–Venue	Original	Dark1	Dark2	Dark3
EmbodiedOcc	2025-ICCV	48.45 / 38.87	45.02 / 35.42	37.32 / 28.86	25.40 / 18.48
EmbodiedOcc++	2025-ACM MM	52.61 / 42.94	48.48 / 38.73	39.85 / 30.33	25.91 / 18.85
GPOcc	2026-CVPR	59.25 / 53.91	57.49 / 52.45	51.89 / 46.53	40.75 / 35.20
MLA	-	58.13 / 53.20	58.14 / 53.87	56.69 / 52.65	53.71 / 49.27
DarkOcc	-	**59.75 / 54.51**	**58.83 / 54.22**	**58.41 / 53.41**	**56.62 / 50.65**

Table 2 shows that DarkOcc improves over MLA by 1.62 / 1.31 points on Original and 2.91 / 1.38 on Dark3. The Original-to-Dark3 Geo-IoU gap decreases from 4.42 to 3.13 points. Gains beyond mixed-light training therefore persist after scene accumulation. Under Dark3, the complete system exceeds the original GPOcc by 15.87 / 15.45 points.

4.4 COMPARISON WITH IMAGE ENHANCEMENT

On Occ-ScanNet-mini, we compare task-level adaptation with image enhancement followed by a fixed GPOcc predictor. The enhancement methods are LCDPNet (Wang et al., 2022), CSEC (Li et al., 2024), HVI-CIDNet (Yan et al., 2025), and Multinex (Brateanu et al., 2026). In Table 3, each enhancer name denotes preprocessing followed by GPOcc; all pipelines receive the same images across Original and Dark1–Dark3.

Table 3: Image-enhancement comparison on the Occ-ScanNet-mini dataset.

Method	Year–Venue	Original	Dark1	Dark2	Dark3
LCDPNet → GPOcc	2022-ECCV	61.83 / 54.60	58.76 / 51.79	50.92 / 43.92	34.05 / 27.21
CSEC → GPOcc	2024-CVPR	61.49 / 54.23	58.20 / 51.20	49.99 / 42.90	33.81 / 26.78
HVI-CIDNet → GPOcc	2025-CVPR	62.22 / 54.91	61.32 / 54.30	58.96 / 51.54	48.61 / 40.68
Multinex → GPOcc	2026-CVPR	61.97 / 54.58	59.71 / 52.58	52.68 / 45.69	36.76 / 29.62
DarkOcc	-	64.20 / 57.35	64.10 / 57.26	63.67 / 56.48	61.96 / 53.73

Compared with GPOcc in Table 1, all enhancement pipelines in Table 3 reduce Original accuracy. HVI-CIDNet performs best among the enhancers across all four conditions and substantially improves Dark2 and Dark3 predictions. Nevertheless, DarkOcc exceeds it by 13.35 Geo-IoU / 13.05 mIoU points on Dark3 while improving normal-light accuracy, supporting adaptation within the occupancy predictor.

4.5 ABLATION STUDIES

Table 4 compares three cumulative configurations on Occ-ScanNet-mini under the same evaluation protocol. Each configuration is trained as a complete model for 30 epochs using Original and Dark1–Dark3: MLA applies the unchanged GPOcc algorithm to mixed illumination data; adding MFA yields the illumination-adapted base (IAB), and adding GOC yields DarkOcc. The original GPOcc in Table 1 provides the Original-only training reference.

Multi-level feature adaptation. Adding MFA improves over MLA by 2.40 / 2.06 points on Original and 2.63 / 2.09 on Dark3. The Original-to-Dark3 Geo-IoU gap changes from 2.68 to 2.45 points. MFA thus improves accuracy under both conditions while retaining the small degradation gap achieved by mixed-light training.

Table 4: Ablation results on Occ-ScanNet-mini.

Method	Original	Dark1	Dark2	Dark3
MLA	59.07 / 54.55	58.76 / 54.01	58.46 / 53.55	56.39 / 50.82
+ MFA	61.47 / 56.61	61.37 / 56.52	60.90 / 55.73	59.02 / 52.91
+ GOC	**64.20 / 57.35**	**64.10 / 57.26**	**63.67 / 56.48**	**61.96 / 53.73**

Geometry-Guided Occupancy Completion. Adding GOC to IAB improves accuracy by 2.94 / 0.82 points on Dark3 and 2.73 / 0.74 on Original. Completion therefore complements feature adaptation under both conditions, consistent with low light exacerbating occupancy incompleteness already present under normal illumination. These cumulative ablations assess MFA and GOC as complete modules.

Generalization to unseen degradations. We also evaluate spatially varying attenuation (U1) and altered noise and color responses (U2), both excluded from training. DarkOcc improves over MLA by 3.87 / 1.40 points on U1 and 3.91 / 1.34 points on U2. The 95% scene-bootstrap confidence intervals for gains over both MLA and IAB are positive for both metrics, as reported in Appendix D. These tests hold scene content fixed and vary the degradation process, supporting generalization beyond the synthetic transformations used for training.

Inference efficiency. On a single RTX 8000, DarkOcc averages 354.32 ms per frame versus 322.67 ms for GPOcc, adding modest inference cost. Appendix F provides the timing protocol and detailed results.

4.6 Evaluation on Real Indoor Data

Table 5 compares six methods on the real-data evaluation set defined in Appendix E.3, with scores pooled separately for normal and low light. DarkOcc achieves the highest scores under both conditions. Relative to GPOcc, the gains are 7.38 / 2.87 points under normal light and 13.86 / 7.69 under low light. Under low light, it also exceeds MLA by 9.76 / 4.05 points and IAB by 5.15 / 1.69; the normal-light gains over MLA are 9.30 / 3.46. Per-scene and class-wise results are provided in Appendices E.3 and G.2.

Table 5: Prediction performance on real indoor data.

	Normal light		Low light	
Method	Geo-IoU	mIoU	Geo-IoU	mIoU
EmbodiedOcc	25.09	15.54	14.40	4.41
FLM-Occ	22.11	14.31	11.56	4.12
GPOcc	39.00	22.66	15.26	7.87
MLA	37.08	22.07	19.36	11.51
IAB	42.86	23.93	23.97	13.87
DarkOcc	**46.38**	**25.53**	**29.12**	**15.56**

5 Conclusion

We presented DarkOcc, an indoor occupancy prediction framework combining Multi-Level Feature Adaptation with Geometry-Guided Occupancy Completion. The framework addresses degraded visual representations and occupancy incompleteness amplified by low light. Experiments on monocular and embodied benchmarks show gains beyond mixed-light training under both normal and low illumination. Evaluation on our newly constructed real-world benchmark also shows improved accuracy on captured low-light images. These findings highlight the complementary roles of representation adaptation and volumetric completion in illumination-robust occupancy prediction.

AI USE STATEMENT

We used generative AI tools to assist with literature discovery, manuscript drafting and editing, feedback on experimental methodology, and interpretation of results. AI tools also assisted in inspecting data and 3D outputs and in identifying and correcting geometric and semantic annotation issues during occupancy ground-truth construction. The annotations were constructed through the algorithmic pipeline described in this paper, with AI assistance in inspection and correction. The authors are responsible for validating all AI-assisted contributions and for the final content, data, and results.

REFERENCES

Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Alexandru Brateanu, Tingting Mu, Codruta O. Ancuti, and Cosmin Ancuti. Multinex: Lightweight low-light image enhancement via multi-prior Retinex. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. https://arxiv.org/abs/2604.10359.

Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage Retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. https://arxiv.org/abs/2303.06705.

Guangcheng Chen, Lihuang Fang, Huaqi Tao, Yicheng He, Li He, and Hong Zhang. FLM-Occ: Feed-forward likelihood maximization for efficient indoor occupancy prediction. In *European Conference on Computer Vision*, 2026. https://arxiv.org/abs/2606.21373.

Haoming Chen, Lichen Yuan, Tianfang Sun, Jingyu Gong, Xin Tan, Zhizhong Zhang, and Yuan Xie. YouTube-Occ: Learning indoor 3D semantic occupancy prediction from YouTube videos. arXiv preprint arXiv:2506.18266, 2025. https://arxiv.org/abs/2506.18266.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839, 2017.

Daniel Feijoo, Juan C. Benito, Alvaro Garcia, and Marcos V. Conde. DarkIR: Robust low-light image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. https://arxiv.org/abs/2412.13443.

Yiran Guo, Simone Mentasti, Xiaofeng Jin, Matteo Frosi, and Matteo Matteucci. SGR-OCC: Evolving monocular priors for embodied 3D occupancy prediction via soft-gating lifting and semantic-adaptive geometric refinement. arXiv preprint arXiv:2603.14076, 2026. https://arxiv.org/abs/2603.14076.

Zeyu Jiang, Changqing Zhou, Xingxing Zuo, and Changhao Chen. FreeOcc: Training-free embodied open-vocabulary occupancy prediction. arXiv preprint arXiv:2604.28115, 2026. https://arxiv.org/abs/2604.28115.

Jianing Li, Ming Lu, Hao Wang, Chenyang Gu, Wenzhao Zheng, Li Du, and Shanghang Zhang. SliceOcc: Indoor 3D semantic occupancy prediction with vertical slice representation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2025a. https://arxiv.org/abs/2501.16684.

Xiang Li, Yupeng Zheng, Pengfei Li, Yilun Chen, Ya-Qin Zhang, and Wenchao Ding. Enhancing indoor occupancy prediction via sparse query-based multi-level consistent knowledge distillation. *IEEE Robotics and Automation Letters*, 10(11):11690–11697, 2025b. https://doi.org/10.1109/LRA.2025.3615532.

Yiyu Li, Ke Xu, Gerhard Petrus Hancke, and Rynson W. H. Lau. Color shift estimation-and-correction for image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ze Li, Feng Zhang, Xiatian Zhu, Meng Zhang, Yanghong Zhou, and P. Y. Mok. Robust low-light scene restoration via illumination transition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025c. https://arxiv.org/abs/2507.03976.

Aoyu Liu, Zhen Liu, Ziyi Wang, Dian Chen, Bing Zeng, and Shuaicheng Liu. ExpoCM: Exposure-aware one-step generative single-image HDR reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. https://arxiv.org/abs/2605.02464.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. https://openreview.net/forum?id=Bkg6RiCqY7.

Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. NeRF in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16190–16199, 2022.

Mingyang Ou, Haojin Li, Yifeng Zhang, Ke Niu, Zhongxi Qiu, Heng Li, and Jiang Liu. DeLight-Mono: Enhancing self-supervised monocular depth estimation in endoscopy by decoupling uneven illumination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. https://arxiv.org/abs/2511.20058.

Rui Qian, Haozhi Cao, Tianchen Deng, Shenghai Yuan, and Lihua Xie. SplatSSC: Decoupled depth-guided Gaussian splatting for semantic scene completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(10):8520–8528, 2026. https://doi.org/10.1609/aaai.v40i10.37803.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, 2021.

Yuan Sun, Julio Contreras, and Jorge Ortiz. Dynamic focused masking for autoregressive embodied occupancy prediction. In *Advances in Neural Information Processing Systems*, 2025. https://proceedings.nips.cc/paper_files/paper/2025/hash/ec3d49763c653ad7c8d587f52220c129-Abstract-Conference.html.

Hao Wang, Xiaobao Wei, Xiaoan Zhang, Jianing Li, Chengyu Bai, Ying Li, Ming Lu, Wenzhao Zheng, and Shanghang Zhang. EmbodiedOcc++: Boosting embodied 3D occupancy prediction with plane regularization and uncertainty sampler. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025a.

Haoyuan Wang, Ke Xu, and Rynson W. H. Lau. Local color distributions prior for image enhancement. In *European Conference on Computer Vision*, 2022.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025b.

Ruoyu Wang, Yong Liu, Sheng Tao, Yuhang Lin, and Yukai Ma. VEOcc: Voxel-centric online semantic occupancy prediction for embodied scene understanding. arXiv preprint arXiv:2605.25059, 2026. https://arxiv.org/abs/2605.25059.

Yue Wen, Liang Song, and Hesheng Wang. SplatBright: Generalizable low-light scene reconstruction from sparse views via physically-guided Gaussian enhancement. arXiv preprint arXiv:2512.18655, 2025. https://arxiv.org/abs/2512.18655.

Yuan Wu, Zhiqiang Yan, Yigong Zhang, Xiang Li, and Jian Yang. See through the dark: Learning illumination-affined representations for nighttime occupancy prediction. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Yuqi Wu, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Jie Zhou, and Jiwen Lu. EmbodiedOcc: Embodied 3D occupancy prediction for vision-based online scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Rui Xu, Yuzhen Niu, Yuezhou Li, Huangbiao Xu, Wenxi Liu, and Yuzhong Chen. URWKV: Unified RWKV model with multi-state perspective for low-light image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. https://arxiv.org/abs/2505.23068.

Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. SNR-aware low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. https://openaccess.thecvf.com/content/CVPR2022/html/Xu_SNR-Aware_Low-Light_Image_Enhancement_CVPR_2022_paper.html.

Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. HVI: A new color space for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5678–5687, 2025.

Zhen Yao and Mooi Choo Chuah. Event-guided low-light video semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. https://openaccess.thecvf.com/content/WACV2025/html/Yao_Event-Guided_Low-Light_Video_Semantic_Segmentation_WACV_2025_paper.html.

Hongxiao Yu, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Monocular occupancy prediction for scalable indoor scenes. In *European Conference on Computer Vision*, pp. 38–54, 2024.

Longjian Zeng, Zunjie Zhu, Rongfeng Lu, Ming Lu, Bolun Zheng, Chenggang Yan, and Anke Xue. DepthDark: Robust monocular depth estimation for low-light environments. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. https://arxiv.org/abs/2507.18243.

Tianyi Zhang, Kaining Huang, Weiming Zhi, and Matthew Johnson-Roberson. DarkGS: Learning neural illumination and 3D Gaussians relighting for robotic exploration in the dark. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024. https://arxiv.org/abs/2403.10814.

Changqing Zhou, Yueru Luo, and Changhao Chen. Generalizing visual geometry priors to sparse Gaussian occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28578–28587, 2026a.

Changqing Zhou, Yueru Luo, Han Zhang, Zeyu Jiang, and Changhao Chen. Monocular open vocabulary occupancy prediction for indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b. https://arxiv.org/abs/2602.22667.

Changqing Zhou, Yueru Luo, Yulan Guo, Bing Wang, Jie Qin, and Changhao Chen. GPOcc++: Unified sparse Gaussian occupancy prediction with visual geometry priors. arXiv preprint arXiv:2607.13481, 2026c. https://arxiv.org/abs/2607.13481.

A Implementation and Training Details

A.1 Architecture and Optimization

Table 6 collects the numerical settings for the DarkOcc model. MFA operates on the encoder patch grid, whereas image descriptors and depth cues are computed on the higher-resolution ray grid. The base initializes 66,304 Gaussians before opacity pruning. MFA adds 550,272 parameters. The GOC branch comprises the image adapter, width predictor, and voxel repair network listed in Table 8. Tables 7 and 8 summarize parameter groups optimized within the same training run.

Table 6: Architecture and optimization settings.

Setting	Value
Input / patch grid / ray grid	392×518 / 28×37 / 56×74
Voxel grid / side length	$60 \times 60 \times 36$ / 0.08 m
VGGT adapter interfaces / bottleneck	4, 11, 17, 23 / 32 channels
Per-image feature residual cap ρ_F	0.03
Base Gaussian samples per ray N_G	16
Base range / extent	$[0, 8]$ m / $[0, 5]$ m
Image descriptor	32 channels
Repair base input / hidden state	58 / 32 channels
Learned width τ / initial width	$(0.08, 0.40)$ m / 0.16 m
Squared normalized depth-offset cap	80
Numerical stabilizer ϵ	10^{-12}
Repair CE occupied/free weights	2 / 1
Repair weights $\lambda_{\mathrm{CE}}, \lambda_{\mathrm{Lov}}$	1 / 1
AdamW weight decay / gradient clipping	0.01 / 1.00

Joint training protocol. We use 5,504 training and 2,376 test frames for Occ-ScanNet-mini, and 45,610 training and 19,705 test frames for full Occ-ScanNet. MLA, IAB, and DarkOcc are each trained for 30 epochs on the corresponding training split using mixed illumination data covering Original and Dark1–Dark3. Initialization and sampling use fixed seeds.

Table 7: Backbone and MFA optimizer settings.

Group	Peak LR	Terminal LR
Feature adapters	2×10^{-4}	10^{-6}
Task heads	10^{-4}	10^{-6}
VGGT	10^{-5}	10^{-7}

Supervision. Following GPOcc (Zhou et al., 2026a), each view receives focal classification, Lovász overlap (Berman et al., 2018), semantic and geometric scaling, and camera-depth Huber supervision. The respective weights are 100, 1, 1, 1, and 0.20. Focal loss uses focusing parameter 2 and balancing factor 0.25, with inverse-log class-frequency weighting. The Huber transition is 1 m. For the GOC branch, cross-entropy is normalized by the number of editable known voxels, with weights 2 for occupied targets and 1 for free space. The occupied-class Lovász term uses all valid supervised voxels: editable voxels use the repaired distribution, while protected voxels use the base decision. Losses are averaged across views and normalized over valid supervision elements.

Joint optimization and parameter groups. The backbone, MFA, and GOC are optimized within the same 30-epoch training run. The backbone receives task supervision, while completion losses supervise the GOC branch. Optimization uses AdamW with learning-rate warmup, cosine decay, FP32, and gradient clipping at 1.0.

Prediction path. The final model consists of the base with MFA and its GOC branch. The GOC branch uses the original depth and voxel scores predicted by the occupancy base throughout. Local predictions provide single-frame occupancy outputs; the embodied prediction protocol accumulates their signed repair residuals alongside the base Gaussian memory, as detailed in Appendix B.1.

Table 8: GOC parameter counts and learning rates.

Group	Parameters	Peak LR
Image adapter	8,672	2×10^{-4}
Geometry-Guided Occupancy Completion network	10,668	2×10^{-4}
Width predictor	33	5×10^{-5}

A.2 Feature Modulation

The feature gates are computed from detached RGB statistics. Luminance uses channel weights $(0.2126, 0.7152, 0.0722)$. Within each patch, the local response combines low mean luminance and the fraction of pixels below 0.25, with weights 0.75 and 0.25. A contrast contribution, scaled by 0.10, increases the response in locally structured regions. The global response applies a smoothstep to mean-image luminance between approximately 0.30 and 0.35. Multiplying these responses modulates the spatial residual in Eq. (2); the original image remains the encoder input.

A.3 Voxel Features for Occupancy Completion

The image descriptor is sampled at each voxel's projection and modulated by proximity to the original base depth as in Eq. (3). The width predictor maps its output through a sigmoid to the interval in Table 6; its width controls depth influence rather than calibrated uncertainty. For numerical stability, the squared normalized depth offset is capped at 80.

The projected voxel input contains 32 image channels, 12 base-score channels, and 14 depth/projection channels. These encode camera depth, sampled base depth, surface offsets, near-surface strength, illumination, extent, image position, validity, and the base-free indicator. Depth and extent are normalized by their respective maximum ranges. The local and coarse repair branches each contain two depthwise 3D residual blocks, with coarse features upsampled before the output projection.

A valid projection lies inside the image with positive finite camera depth; invalid features are zeroed. Outside this region, the full base score vector is copied unchanged.

B Dataset Splits and Evaluation Protocol

Partitions and scope. We follow the official Occ-ScanNet-mini split, using all 5,504 training frames for training and the 2,376 test frames for evaluation. The full-data experiment uses 45,610 training frames and 19,705 test frames in each of the four conditions. Embodied evaluation on the mini split contains 16 sequences and 480 frames per condition. The same test inputs, labels, and metric definitions are used for every method.

Monocular protocols. DiScene uses its distilled depth-prior model, including for mini evaluation, with the official sampling fallback. Stochastic test-time color transforms and functional dropout are disabled for FLM-Occ. RGB-derived priors are recomputed for every degraded input.

All synthetic comparisons use a unified evaluation protocol with identical test frames, target labels, evaluation masks, and metric computation. The Original and Dark3 entries in Table 11 use the same results as the main ablation table; U1 and U2 evaluate the corresponding models on unseen synthetic degradations.

B.1 Gaussian and Residual Fusion for Embodied Occupancy Prediction

Embodied occupancy prediction combines the adapted occupancy base with the GOC branch. Each frame produces the original Gaussian bundle and a signed valid-class score residual between repaired and local base predictions. The Gaussian memory follows GPOcc's incremental update (Zhou et al., 2026a), while residuals are accumulated separately so that voxel corrections enter the global output.

Gaussians are transformed into world coordinates and merged within 0.02 m, using an opacity threshold of 0.01 and an old-frame weight of 0.10. Global queries are mapped to the nearest local voxel on the 8 cm grid. Residuals are averaged across all valid observations, including zero-residual

views, and have zero value where unobserved. The mean is added with unit gain only where the current global Gaussian base predicts free space. Repaired scores do not feed back into Gaussian memory. Scene memory is reset between sequences, and model parameters remain fixed during embodied inference.

All embodied occupancy prediction methods are evaluated on the same 16 sequences and the same global target domain after the thirtieth frame. The evaluation uses identical observation masks and label handling across methods. We pool confusion counts across scenes before computing Geo-IoU and mIoU.

C Synthetic Low-Light Generation

We generate synthetic low-light observations by changing image formation while preserving camera parameters and occupancy labels. Motivated by the role of signal-dependent noise in low-light imaging (Mildenhall et al., 2022), the transformation combines exposure attenuation, a gamma adjustment, and signal-dependent, read, and row-correlated noise. For normalized linear RGB x, the attenuated signal is $\mu_s = (g_s x)^{\gamma_s}$ and the noisy output is

$$\widetilde{x}_s = \mathrm{clip}(\mu_s + \sqrt{\alpha_s \mu_s} \odot \varepsilon_s + \sigma_{\mathrm{read},s}\varepsilon_r + \sigma_{\mathrm{row},s}\varepsilon_{\mathrm{row}}, 0, 1)\,. \tag{5}$$

The Dark1–Dark3 protocol includes intermediate 8-bit quantization in linear RGB; it therefore evaluates robustness to the combined effects of attenuation, noise, and quantization. The noise variables are standard normal. Signal-dependent and read noise vary per pixel and channel, whereas a row-noise sample is shared across columns and channels. Table 9 defines the three severity levels. Outputs are quantized to 8 bits in linear space, converted to sRGB, and quantized again when saved as PNG. Per-image noise realizations are shared across severity levels and scaled by their corresponding amplitudes. Fixed seeds and cached images give all methods identical test inputs.

Table 9: Synthetic low-light parameters.

Parameter	Dark1	Dark2	Dark3
Exposure gain g_s	0.60	0.35	0.20
Gamma γ_s	1.08	1.20	1.35
Signal-dependent noise α_s	4.00×10^{-3}	7.00×10^{-3}	1.10×10^{-2}
Read-noise deviation $\sigma_{\mathrm{read},s}$	2.00×10^{-3}	4.00×10^{-3}	6.00×10^{-3}
Row-noise deviation $\sigma_{\mathrm{row},s}$	1.00×10^{-3}	2.00×10^{-3}	3.00×10^{-3}

D Generalization to Unseen Synthetic Degradations

D.1 Alternative Image Formation

We evaluate the trained models under two transformations excluded from training, using the same 2,376 source frames and common evaluation protocol. U1 introduces spatially varying attenuation through a smooth positive field with random horizontal and vertical phases, normalized to the mean exposure gain in Table 10. This changes where illumination is concentrated without applying the original gamma and noise recipe.

U2 combines Poisson shot noise with Laplace read noise, column-correlated Gaussian noise, and color variation. We sample one of the two RGB gain vectors in Table 10, normalize it to unit luminance, and apply it to the attenuated linear signal μ. The result is

$$y_2 = \mathrm{Poisson}(N_p\mu)/N_p + \varepsilon_{\mathrm{read}} + \varepsilon_{\mathrm{col}}. \tag{6}$$

The read-noise scale and column-noise standard deviation are given in the table. Column noise is shared across rows and color channels. This condition tests a joint change in noise distribution, spatial correlation, and color response; it does not isolate the effect of Poisson noise alone. Both transformations clip the linear output, convert it to sRGB, and save 8-bit PNGs without intermediate linear quantization. Their difficulty is not calibrated to match Dark3, so comparisons are made between methods within each condition.

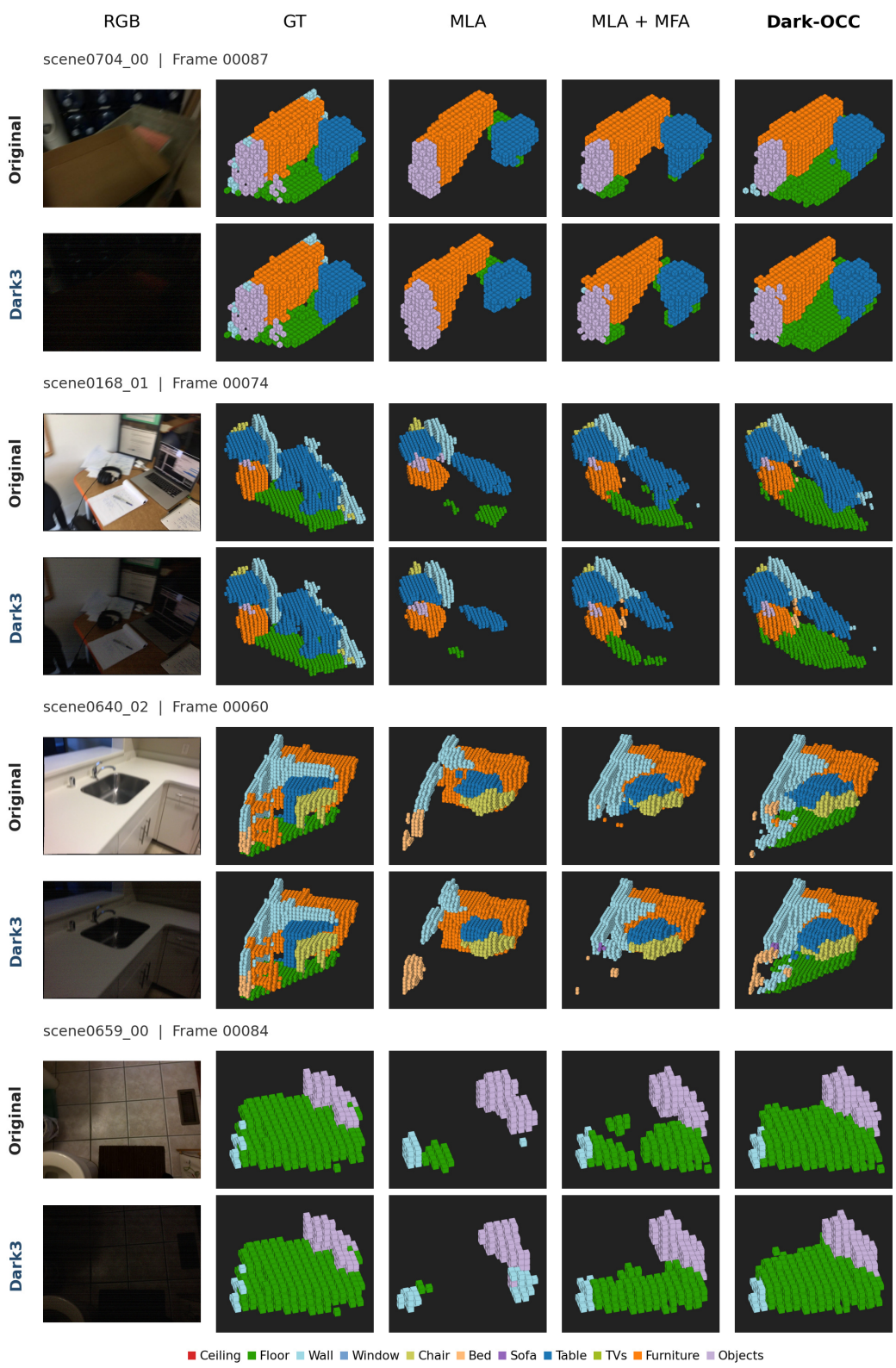


Figure 4: Qualitative comparison on Occ-ScanNet-mini (part 1 of 2). Each scene shows paired Original and Dark3 inputs, ground truth, MLA, MLA+MFA, and DarkOcc. All occupancy views within each scene share the same camera and scale.

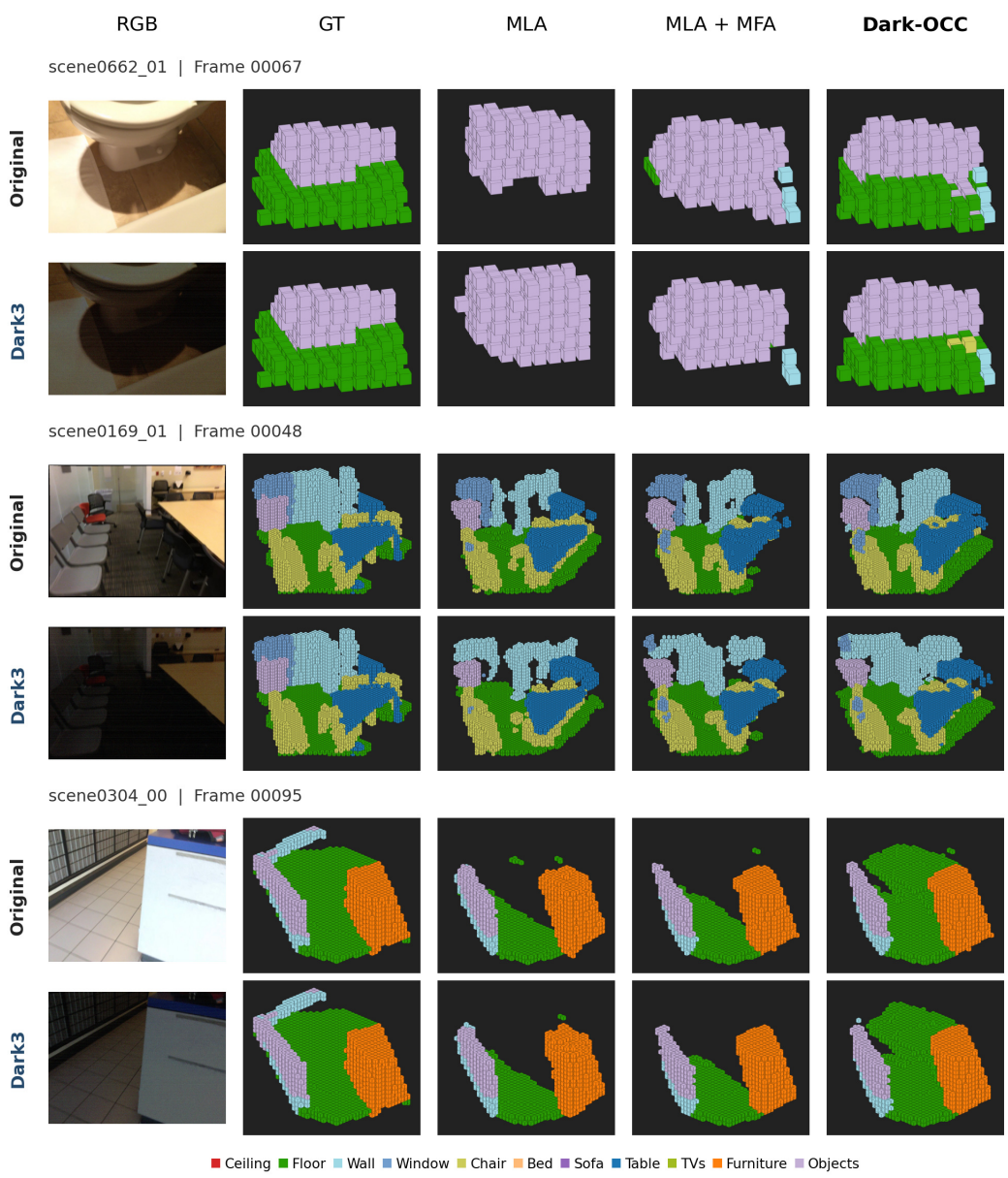


Figure 5: Qualitative comparison on Occ-ScanNet-mini (part 2 of 2), continuing Fig. 4. The panel order and display scale are unchanged.

Table 10: Parameters for unseen synthetic degradations.

Condition	Parameter	Value
U1	Mean exposure gain g_U	0.06
U1	Spatial amplitudes a_x, a_y	0.70 / 0.40
U2	Exposure gain g_U	0.06
U2	Photon-count scale N_p	256
U2	Read-noise scale b_r	1.50×10^{-3}
U2	Column-noise deviation σ_c	1.00×10^{-3}
U2	RGB gains before normalization	$(1.40, 1, 0.70)$ or $(0.75, 1, 1.35)$

Table 11: Prediction performance under unseen synthetic degradations.

Method	Original	Dark3	U1	U2
MLA	59.07 / 54.55	56.39 / 50.82	57.41 / 52.86	57.50 / 52.75
IAB	61.47 / 56.61	59.02 / 52.91	58.18 / 53.26	58.26 / 53.15
DarkOcc	**64.20 / 57.35**	**61.96 / 53.73**	**61.28 / 54.26**	**61.41 / 54.09**

D.2 Results

Table 11 compares the same trained models under two transformations excluded from training. On U1, DarkOcc achieves 61.28 / 54.26, exceeding MLA by 3.87 / 1.40 points and IAB by 3.10 / 1.00 points. On U2, it reaches 61.41 / 54.09, with gains of 3.91 / 1.34 points over MLA and 3.15 / 0.94 points over IAB. The gains over both references show that the complete method improves on mixed-light training and feature adaptation under each unseen degradation. Scene content is held fixed, so the experiment evaluates transfer across degradation processes on the same test scenes.

We estimate evaluation uncertainty through 2,000 paired bootstrap resamples of the 80 scenes, summing confusion matrices before computing each resampled score. The 95% percentile intervals in Table 12 remain positive for both metrics relative to augmentation and the adapted base. They describe variability over the evaluated scenes, not training-seed variability or performance on an independent population of scenes.

Table 12: Confidence intervals for unseen-degradation gains.

Condition	Reference method	Geo-IoU gain	mIoU gain
U1	IAB	$[2.43, 3.92]$	$[0.75, 1.26]$
U1	MLA	$[3.29, 4.56]$	$[0.63, 2.42]$
U2	IAB	$[2.46, 3.98]$	$[0.71, 1.21]$
U2	MLA	$[3.31, 4.60]$	$[0.66, 2.20]$

E Real Indoor Data Collection and Annotation

E.1 Data and Evaluation Scope

Table 13 summarizes the evaluation set from 19 recordings in 10 real scenes. Normal-light evaluation covers nine scenes and low-light evaluation covers all ten; Room 02 has no normal-light recording. Counts refer to the evaluated frames. RGB remains unaltered for inference, while geometric observations and camera calibration are used offline to construct reference labels.

Table 13: Real-data evaluation statistics.

Illumination	Scenes	Recordings	Evaluated frames
Normal light	9	9	3,222
Low light	10	10	4,986
Total	10	19	8,208

E.2 Reference Geometry and Completion

Trajectory and surface reconstruction. The construction pipeline first associates the recorded streams and recovers metric depth using the sensor calibration. Offline trajectory estimation combines depth-based registration, available motion priors, and loop-closure optimization. Candidate trajectories are assessed using depth reprojection consistency and reconstruction alignment rather than occupancy-model scores. Registered depth frames are fused into a high-resolution TSDF/mesh before conversion to the coarser occupancy grid. We refer to these trajectories as optimized reference poses because no independent motion-capture ground truth is available.

Semantic annotation and completion. Multi-view image masks are projected onto the reconstructed scene and associated into semantic instances. Geometric plane fitting provides room-layout candidates; simple primitives and CAD alignment provide object-completion candidates where direct surface measurements are incomplete. Candidate geometry is checked against multi-view depth and RGB projections and is admitted to the reference only after review. Measured surfaces, reviewed layout/primitive completion, and reviewed CAD completion retain separate provenance. Automatic proposals awaiting review are excluded from the valid annotation domain. This construction combines observed geometry with explicit, inspectable completion rather than treating network predictions as ground truth.

Voxel labels and evaluation frames. Reference geometry follows the Occ-ScanNet coordinate and semantic conventions. Evaluation retains each recording's documented target-grid coverage, including grids with 36 or 72 vertical layers, and its recorded scale settings. These include unit scale and method-specific calibrated scale factors. Rays terminating at valid measured surfaces identify observed free space; unsupported regions remain unknown unless supported by reviewed geometry. Geometry and semantic validity are stored separately, and geometry displayed for annotation review is not automatically included in either scoring domain. Models receive RGB and camera parameters, without sensor depth. The evaluation protocol is described below.

E.3 Per-scene Evaluation

The normal- and low-light recordings follow independent trajectories rather than frame-aligned illumination pairs. All methods use the same evaluation frames within each recording. The evaluation set was assembled with reference to observed model performance and temporal coverage; its scores therefore characterize this set under the recorded grid and scale settings rather than an independent test of generalization. Differences between lighting conditions also reflect changes in viewpoints, visible content, and scene coverage.

Table 14 compares GPOcc, MLA, the illumination-adapted base (IAB, i.e., MLA with MFA), and DarkOcc. Each cell reports Geo-IoU / mIoU in percent. Scores are computed from pooled voxel-level counts rather than averaged frame scores. For each recording, semantic IoU is averaged over a category set fixed from its full evaluable ground truth and shared by all methods. The category set can differ across recordings, so cross-scene mIoU differences also reflect category composition.

Table 14: Per-scene prediction performance on real data.

Scene	Frames	GPOcc	MLA	IAB	DarkOcc
Normal light					
Room 01	108	32.53 / 9.07	25.44 / 7.06	32.70 / 9.11	**37.22 / 10.55**
Bedroom A	719	40.27 / 16.91	40.08 / 16.74	44.99 / 18.02	**48.30 / 19.18**
Bedroom B	324	36.32 / 14.63	35.09 / 14.63	43.07 / 17.34	**43.71 / 17.72**
Common area	232	47.17 / 27.83	40.07 / 22.94	48.89 / 26.55	**52.50 / 28.74**
Office	475	40.34 / 28.29	40.14 / 28.18	44.59 / 30.24	**46.50 / 31.31**
Bedroom C	321	40.26 / 18.99	32.32 / 17.32	42.85 / 20.53	**44.00 / 22.06**
Living room	295	41.88 / 25.97	39.79 / 25.51	46.87 / 28.09	**49.55 / 29.31**
Meeting room	385	36.43 / 16.64	32.83 / 16.32	38.21 / 18.69	**45.54 / 22.38**
Laboratory	363	34.48 / 12.50	32.72 / 11.88	38.59 / 13.54	**43.18 / 15.29**
Low light					
Room 01	816	12.47 / 3.70	18.96 / 5.43	25.37 / 7.98	**33.96 / 9.69**
Room 02	834	22.21 / 4.72	22.43 / 5.46	30.52 / 7.26	**38.22 / 8.62**
Bedroom A	822	25.00 / 10.93	28.09 / 12.59	35.16 / 16.68	**39.05 / 18.16**
Bedroom B	153	18.55 / 4.99	25.35 / 8.72	26.05 / 9.03	**26.70 / 9.13**
Common area	390	6.06 / 3.93	11.29 / 8.53	11.70 / 8.97	**13.03 / 9.73**
Office	558	18.98 / 9.87	25.58 / 15.03	30.35 / 17.43	**32.59 / 18.47**
Bedroom C	133	11.16 / 6.10	15.24 / 9.64	25.23 / 11.42	**25.67 / 12.17**
Living room	304	19.88 / 13.03	22.35 / 13.04	26.38 / 15.85	**32.82 / 17.83**
Meeting room	717	10.06 / 5.45	13.46 / 8.38	16.44 / 10.24	**22.52 / 13.46**
Laboratory	259	8.94 / 2.75	12.84 / 4.48	14.67 / 5.24	**19.60 / 6.16**

DarkOcc exceeds GPOcc, MLA, and IAB in both metrics for each reported recording, but performance varies substantially across scenes. Under low light, its Geo-IoU ranges from 13.03 in the common area to 39.05 in Bedroom A. The Geo-IoU difference relative to IAB ranges from 0.44 to 8.59 points, indicating that the size of the additional improvement depends on the scene. The low absolute accuracy in difficult recordings also shows that higher aggregate scores do not imply complete reconstruction. These comparisons describe recording-level results, not improvements on every frame or isolated causal effects of individual modules.

E.4 Semantic Accuracy on Ground-Truth Occupied Voxels

To complement IoU, we report the fraction of semantically valid ground-truth occupied voxels assigned the correct class. This accuracy divides the number of correct occupied-class predictions by the total number of valid ground-truth occupied voxels, with counts pooled across the evaluated recordings. It weights voxels equally rather than semantic classes. Both free-space predictions at occupied locations and incorrect semantic labels count as errors; false occupied predictions in ground-truth free space are not directly penalized by this metric.

Table 15: Semantic accuracy on ground-truth occupied voxels.

Method	Normal light (%)	Low light (%)
EmbodiedOcc	21.00	8.31
FLM-Occ	17.72	6.09
GPOcc	32.85	11.00
MLA	30.93	14.79
IAB	35.80	18.55
DarkOcc	**38.76**	**22.48**

Table 15 shows that DarkOcc correctly labels 38.76% of ground-truth occupied voxels under normal light and 22.48% under low light. Relative to MLA, these values are higher by 7.83 and 7.69 percentage points; relative to IAB, the differences are 2.96 and 3.93 points. The results complement the IoU comparisons by showing higher semantic accuracy at truly occupied locations. This metric is not a completion success rate or a direct count of voxels correctly added by GOC, and should be interpreted alongside Geo-IoU and mIoU to account for false occupied predictions.

E.5 Qualitative Results on Real Indoor Data

Comparison with existing methods. Figure 6 compares predictions on real low- and normal-light inputs. In the low-light examples, several baselines produce fragmented volumes or leave substantial portions of the reference structure unrepresented. Mixed-light training alleviates some failures, but large gaps remain in several MLA predictions. DarkOcc recovers more spatially continuous occupied structure and more closely follows the reference layout in these examples. The normal-light rows show that this behavior also extends to well-illuminated inputs. These visual comparisons illustrate the aggregate results in Table 5.

Feature adaptation and occupancy completion. Figure 7 compares MLA, MLA with MFA (the illumination-adapted base, IAB), and the complete DarkOcc model. Feature adaptation substantially improves the recovered layout in the third and fourth examples, yet incomplete regions remain after MFA, including gaps in the second and fourth rows. Adding GOC produces more continuous structure in these regions and in the lower three examples. The progression illustrates the complementary roles of representation adaptation and volumetric completion: stronger features improve the base prediction, while GOC addresses structure that remains missing. Residual holes and semantic discrepancies remain, so the visual improvements should not be interpreted as complete recovery of the reference geometry.

F Inference Cost

We measure inference latency on a single NVIDIA RTX 8000 with batch size 1 and FP32 precision. All configurations use the same 50 test frames from 50 distinct scenes in Occ-ScanNet-mini, evaluated

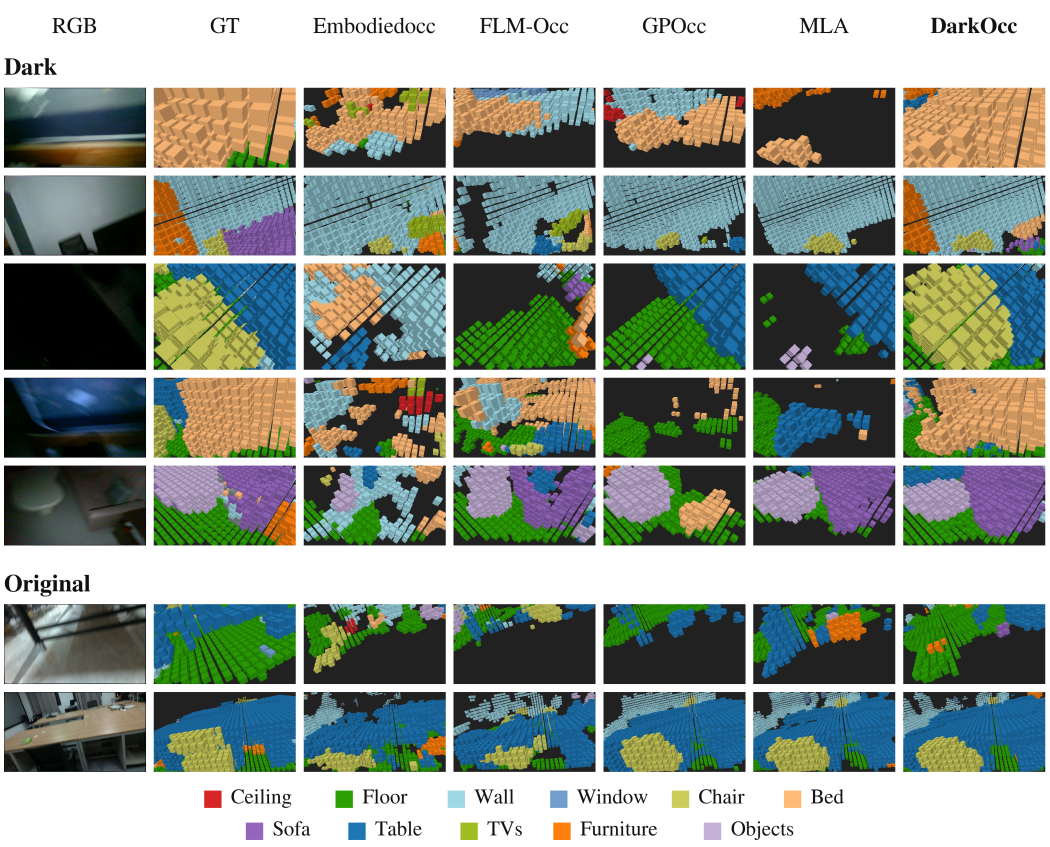


Figure 6: Qualitative comparison on real indoor data. The upper five rows show low-light inputs (Dark), and the lower two show normal-light inputs (Original). Columns present RGB, ground-truth occupancy (GT), and predictions from EmbodiedOcc, FLM-Occ, GPOcc, mixed-light augmentation (MLA), and DarkOcc. Colors indicate semantic categories.

under Original and Dark3. For each condition, we perform 30 warm-up iterations and time each frame three times, synchronizing CUDA before and after each measurement. Timing includes model forward execution and final semantic decoding, but excludes model loading, data loading, preprocessing, input transfer, metric computation, and visualization. Mean latency weights the two illumination conditions equally; equivalent FPS is its reciprocal rather than concurrent throughput.

Table 16: Inference latency on Occ-ScanNet-mini.

Method	Original (ms)	Dark3 (ms)	Mean (ms) ↓	FPS ↑
GPOcc	324.98	320.36	322.67	3.10
IAB	329.71	330.61	330.16	3.03
DarkOcc	346.04	362.60	354.32	2.82

Table 16 reports measured latency for GPOcc, the illumination-adapted base (IAB), and DarkOcc. Paired timing differences on identical frames show that IAB adds 7.49 ms on average over GPOcc, while adding GOC increases latency by a further 24.16 ms. The complete system adds 31.65 ms on average, introducing modest additional inference cost relative to GPOcc. These differences measure net changes in the inference pipeline, rather than isolated execution times of individual modules.

G Class-wise Occupancy Results

We report Geo-IoU, semantic IoU for each of the 11 occupied categories, and mIoU, defined as the unweighted mean of these 11 class-wise IoUs. Category order and colors follow the indoor occupancy

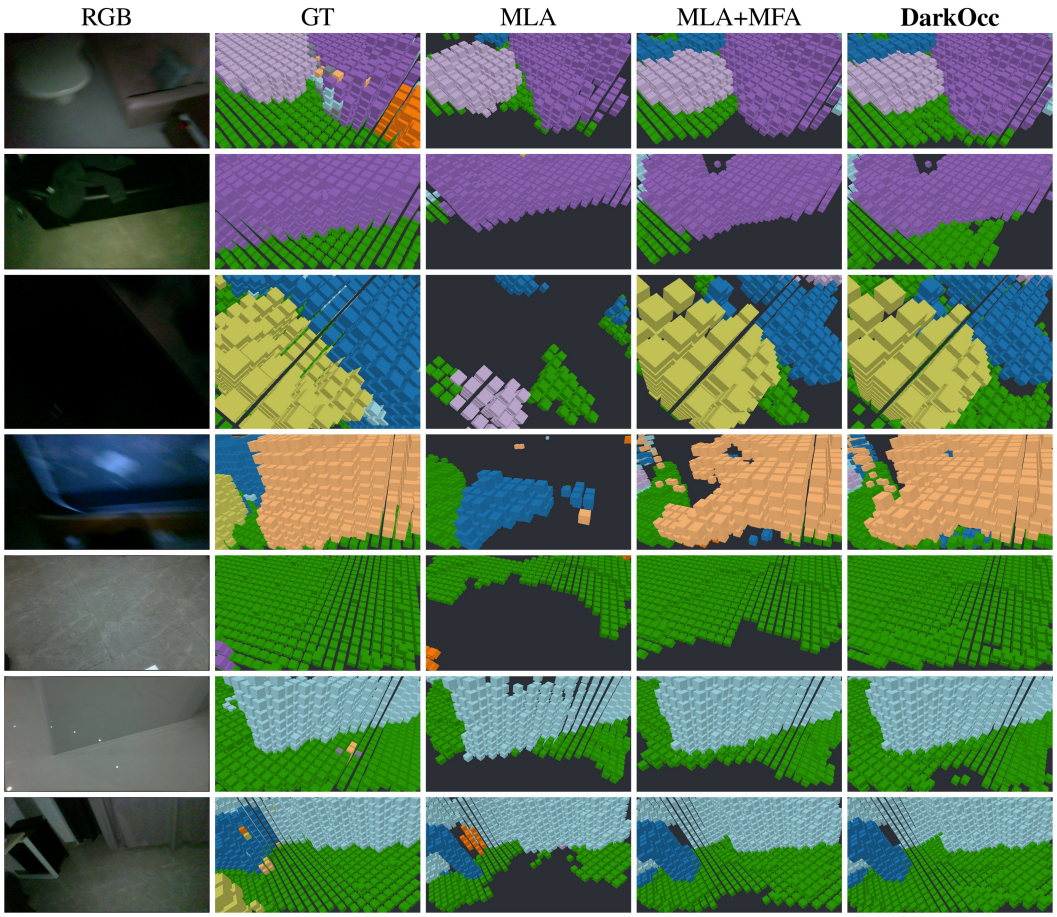


Figure 7: Qualitative component comparison on real indoor data. From left to right: RGB, ground-truth occupancy (GT), MLA, MLA with multi-level feature adaptation (MFA), and DarkOcc with Geometry-Guided Occupancy Completion (GOC). The examples show remaining occupancy incompleteness after feature adaptation and the additional structure recovered by GOC. Semantic colors follow Figure 6.

benchmarks (Wu et al., 2025b; Zhou et al., 2026a). Free space is excluded from the semantic mean. Each category score is computed from pooled intersection and union counts before averaging across classes; it is not an average of per-frame or per-scene IoUs.

G.1 Monocular Prediction on Occ-ScanNet

Table 17 reports class-wise results on Occ-ScanNet under Original and Dark3. DarkOcc achieves the highest class IoU among the reported methods for 10 of the 11 occupied categories under Original and all 11 under Dark3. The aggregate mIoU gains are thus accompanied by improvements across a broad range of semantic categories in both conditions. The aggregate scores match Table 1.

G.2 Real-World Prediction

Table 18 reports class-wise IoU for the real-data evaluation set described in Appendix E.3. Normal- and low-light recordings are pooled separately. All 11 occupied categories occur in each aggregate, so mIoU is their unweighted mean after pooling intersection and union counts. Category order and colors match Table 17; bold indicates the highest displayed value within each lighting condition.

DarkOcc raises aggregate mIoU from 22.07 to 25.53 under normal light and from 11.51 to 15.56 under low light relative to MLA. Compared with IAB, the differences are 1.60 and 1.69 points, respectively. The class-wise profiles show that these aggregate improvements are not uniform across

Table 17: Class-wise prediction performance on Occ-ScanNet.

Method	Geo-IoU	ceiling	floor	wall	window	chair	bed	sofa	table	tvs	furniture	objects	mIoU
Original													
EmbodiedOcc	51.58	37.22	47.82	39.18	29.69	39.55	53.51	57.59	41.85	33.34	50.76	38.84	42.67
SplatSSC	62.84	49.12	59.02	48.28	38.77	47.35	62.39	66.99	49.53	42.60	60.73	45.46	51.84
DiScene	52.02	45.25	50.64	40.46	36.74	42.28	59.66	62.04	45.60	41.14	52.45	42.78	47.19
GPOcc	63.54	55.53	61.43	51.53	46.97	51.32	63.87	70.04	53.62	50.61	62.48	52.39	56.34
FLM-Occ	64.78	56.33	62.77	52.36	44.95	48.82	**67.19**	70.87	54.30	47.52	64.62	52.54	56.57
DarkOcc	**66.38**	**57.39**	**64.79**	**54.18**	**49.96**	**53.41**	66.05	**72.14**	**55.73**	**52.32**	**64.84**	**54.99**	**58.71**
Dark3													
EmbodiedOcc	24.27	9.76	21.22	11.49	6.31	18.56	29.88	24.46	20.12	3.31	25.03	17.18	17.03
SplatSSC	25.69	10.10	16.94	11.85	6.49	21.15	33.15	26.82	22.39	3.36	26.39	18.91	17.96
DiScene	24.20	8.86	19.75	11.93	2.86	13.28	18.31	18.58	13.38	2.08	17.65	9.65	12.39
GPOcc	38.77	21.27	35.50	23.93	16.75	30.44	44.28	41.40	32.91	13.42	39.12	29.49	29.86
FLM-Occ	36.54	20.27	27.53	21.46	15.36	27.83	42.63	37.68	30.10	12.51	34.49	27.33	27.02
DarkOcc	**59.29**	**46.21**	**57.22**	**44.86**	**41.30**	**45.54**	**60.42**	**62.83**	**48.98**	**34.30**	**57.36**	**46.31**	**49.58**

Table 18: Class-wise prediction performance on real data.

Method	Geo-IoU	ceiling	floor	wall	window	chair	bed	sofa	table	tvs	furniture	objects	mIoU
Normal light													
GPOcc	39.00	16.38	20.46	31.27	**2.60**	23.69	30.82	49.57	20.90	1.39	37.47	**14.70**	22.66
MLA	37.08	24.37	15.33	29.49	1.09	24.11	33.02	**49.69**	20.60	0.90	35.72	8.46	22.07
IAB	42.86	**26.04**	19.02	35.76	0.70	25.24	33.10	46.52	22.32	**2.26**	38.19	14.12	23.93
DarkOcc	**46.38**	**26.04**	**30.54**	**36.32**	0.70	**26.22**	**35.58**	46.87	**23.62**	**2.26**	**38.66**	14.08	**25.53**
Low light													
GPOcc	15.26	1.78	9.41	7.53	**2.03**	6.18	17.76	10.26	14.37	0.41	14.59	2.31	7.87
MLA	19.36	**7.36**	9.32	12.74	0.47	11.02	23.28	25.29	14.42	2.09	18.58	1.99	11.51
IAB	23.97	6.53	10.76	16.07	0.69	13.87	25.97	25.31	19.49	**8.14**	22.74	**3.01**	13.87
DarkOcc	**29.12**	6.53	**22.88**	**16.60**	0.68	**14.53**	**28.27**	**26.32**	**21.32**	**8.14**	**22.84**	3.00	**15.56**

categories: some scores remain unchanged or decrease. The table therefore provides a semantic breakdown of the real-data results without attributing the overall gain to any particular category.