# OPEN-VOCABULARY SEGMENTATION IN THE DARK

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Open-vocabulary segmentation has achieved substantial progress in recent years, yet existing methods are mainly developed for well-illuminated images and suffer significant performance degradation under low-light conditions. In this work, we propose OWL-Net, a RAW-assisted framework for open-vocabulary segmentation in the dark. OWL-Net exploits the fine-grained intensity and structural information preserved in RAW data during training, while requiring only RGB input during inference. Specifically, we introduce a RAW EnCoding (REC) module that represents large-range RAW values as the binary bit sequence and extracts RAW features directly at the bit level, enabling critical details within RAW data to be better preserved during feature extraction. Meanwhile, to avoid the reliance on RAW input during inference, we further develop an RGB Representation Enhancement (RRE) module to progressively recover the overwhelmed details and lift the degraded RGB features to RAW-aware space. In addition, we construct LOVS, a real-world Low-light Open-Vocabulary Segmentation dataset containing 1,000 aligned normal-/low-light image pairs, covering 220 semantic categories with 6,949 object-level masks. Extensive experiments on both synthetic low-light benchmarks and LOVS demonstrate that OWL-Net consistently improves different open-vocabulary segmentation frameworks under low-light conditions, without requiring RAW input or explicit image enhancement during inference. Our code and dataset will be released upon acceptance.

## 1 INTRODUCTION

Open-vocabulary segmentation extends conventional segmentation beyond predefined categories, which enables the models to segment arbitrary categories specified by natural-language descriptions (Wu et al., 2024; Zhu & Chen, 2024; Xiao et al., 2026). However, most existing methods are developed for well-illuminated conditions and ignore low-light environments commonly encountered in real-world applications. Consequently, as shown in Fig. 1, the performance of existing methods degrades substantially under severe illumination deficiency, significantly limiting their applicability in practical scenarios.

Performing open-vocabulary segmentation under low-light conditions is particularly challenging (Bi et al., 2025; Wu et al., 2021). Specifically, in extremely dark environments, the overall image response becomes substantially attenuated, making subtle intensity variations and structural details difficult to capture (Jiang et al., 2025; Feijoo et al., 2025). Unlike closed-set segmentation, open-vocabulary segmentation relies on the interaction between visual features and text embeddings to localize categories beyond the predefined label set. However, the limited visual information retained in low-light RGB images makes it difficult to establish reliable correspondences between image regions and textual concepts. As a result, directly performing open-vocabulary segmentation on low-light RGB images often leads to substantial performance degradation.

To address this issue, we propose OWL-Net, which enables robust open-vocabulary segmentation in low-light environments using only RGB images at inference. During training, we introduce RAW data, the minimally processed measurements captured by the CMOS sensor, to complement the limited information of low-light RGB images. RAW data typically have a higher bit depth than standard 8-bit RGB images, thereby preserving finer intensity variations and richer structural details in dark regions (Wang et al., 2024). As shown in the upper-right part of Fig. 1, RAW data exhibit higher channel-averaged entropy than RGB images, indicating richer information. Such information allows the model to exploit object cues that may be severely overwhelmed in low-light RGB images, which



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.

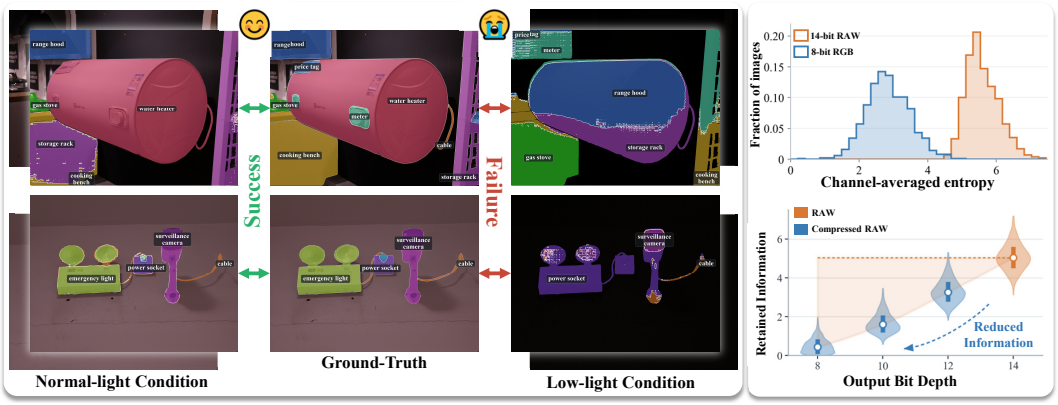



Figure 1: **Left:** Open-vocabulary segmentation performs well under normal illumination but degrades severely in low-light conditions due to insufficient visual cues. **Right:** High-bit-depth RAW data preserve richer information than 8-bit RGB images. However, directly compressing RAW data to a lower bit depth leads to substantial information loss.

could facilitate the establishment of reliable correspondence between image regions and textual concepts. However, owing to its higher bit depth, RAW data typically has a larger numerical range than RGB images. As a result, effectively incorporating RAW data into deep networks to assist the RGB information is non-trivial. A common solution is to directly compress the numerical value of RAW data through normalization and bit quantization (Datta et al., 2023; Li et al., 2025b). However, as depicted in the lower-right part of Fig. 1, such compression causes substantial information loss, discarding the fine-grained details retained in the original RAW data. To address this issue, we introduce a **R**AW **En**Coding (REC) module that represents the high-magnitude RAW value as a binary sequence. REC module estimates the significance of different bit planes based on bit variability and spatial coherence, allowing critical details to be better preserved during feature extraction. In this way, the fine-grained information retained in RAW data can be more effectively incorporated into RGB representations, yielding more discriminative visual features. Therefore, the more reliable correspondences between image regions and textual concepts could be established under low-light conditions, leading to better open-vocabulary segmentation performance.

Meanwhile, unlike previous raw-assisted works that require RAW inputs during inference (Li et al., 2025b; Chen et al., 2023), OWL-Net enables a more flexible inference paradigm without the reliance on extra inputs beyond the RGB image. To achieve this goal, we introduce an RGB Representation Enhancement (RRE) module to convert the RGB features into RAW-aware representations, which are supervised by the real RAW features during training. Through this process, the degraded RGB feature is lifted into the RAW-aware space, recovering visual cues weakened under low-light conditions. By doing so, OWL-Net benefits from the fine-grained information preserved in RAW data even when RAW input is unavailable during inference.

Furthermore, to enable a comprehensive evaluation under real low-light conditions, we construct the Low-light Open-Vocabulary Segmentation (LOVS) dataset. For each scene, we first capture a normal-light image and then recapture the same scene using a 7-stop neutral-density (ND) filter, which attenuates the incoming light by approximately 128×. This acquisition produces paired normal-light and low-light images with pixel-level spatial alignment, allowing accurate annotations to be obtained from the normal-light images. Overall, we collect 1,000 paired normal-/low-light images from diverse real-world environments. Afterwards, we establish the annotation taxonomy from 10 domains covering 220 semantic classes, bringing the 6,949 object-level masks in total.

In summary, our main contributions are as follows:

- We propose OWL-Net, an open-vocabulary segmentation framework designed for low-light conditions. OWL-Net exploits complementary information from RAW data during training to facilitate RGB representations, while requiring only RGB input during inference.
- We develop two plug-and-play modules to exploit RAW information to facilitate the visual feature extraction under low-light conditions. The REC module encodes high-magnitude RAW data while preserving informative bit-level details. The RRE module lifts degraded RGB features into RAW-aware space, enabling flexible RGB-only inference.

- We construct LOVS, a real-world Low-light Open-Vocabulary Segmentation dataset containing 1,000 normal-/low-light image pairs, 220 semantic categories, and 6,949 object-level masks, providing a challenging benchmark for future research.
- Extensive experiments on both synthetic low-light open-vocabulary segmentation benchmarks and the real-world LOVS dataset demonstrate the effectiveness and generality of OWL-Net.

## 2 Related Work

### 2.1 Open-Vocabulary Semantic Segmentation

Open-vocabulary segmentation aims to segment categories beyond a predefined label set based on the text-defined concepts (Wu et al., 2024; Zhu & Chen, 2024). Earlier methods extend conventional segmentation models by aligning pixel-level visual features with textual embeddings for open-vocabulary generalization (Bucher et al., 2019; Xian et al., 2019). More recently, large-scale pretrained foundation models (Kirillov et al., 2023; Radford et al., 2021) have substantially advanced this field by providing stronger visual-semantic representations with more transferable category knowledge. Some approaches adapt vision-language models to facilitate the reliable dense visual-textual correspondences for open-vocabulary segmentation (Xiao et al., 2026; Li et al., 2025a; Cho et al., 2024). The others leverage the segmentation foundation models to generate class-agnostic mask proposals, followed by vision-language models for region classification (Lee et al., 2025; Yuan et al., 2024; Liang et al., 2023). Despite these advances, most existing approaches are developed for well-illuminated conditions, leaving low-light scenarios largely unexplored.

### 2.2 Perception in Low-Light Conditions

Low-light conditions substantially degrade visual information and pose significant challenges to downstream perception tasks. Early approaches often adopt a two-stage pipeline, where low-light images are first enhanced and then processed by downstream perception models (Hashmi et al., 2023; Wang et al., 2024). Although effective, such approaches introduce additional computational overhead and limit inference efficiency. More recent studies therefore focus on direct perception from low-light images, including object detection (Cui et al., 2025; Hashmi et al., 2023), semantic segmentation (Bi et al., 2025; Wei et al., 2023), and tracking (Zhong et al., 2025; Yao et al., 2025), where models are specifically designed to extract robust visual representations under insufficient illumination (Hashmi et al., 2025; Wang et al., 2024). Some methods further exploit auxiliary modalities, such as RAW images (Guo et al., 2025; Li et al., 2025b), infrared images (Ding et al., 2026; Fan et al., 2025), or event data (Li et al., 2024; Baek et al., 2025), to provide complementary information overwhelmed in degraded RGB image. However, most existing methods are developed for predefined categories, and auxiliary modalities are often required during inference. In contrast, we study open-vocabulary segmentation under low-light conditions and leverage RAW data only during training, while maintaining RGB-only inference for flexible deployment.

## 3 Method

### 3.1 Preliminaries

Given an RGB image $I_{low} \in \mathbb{R}^{3\times H\times W}$ and a set of text-defined category concepts $\mathcal{T} = \{t_k\}_{k=1}^{K}$, our goal is to predict a dense semantic map $M$, where each pixel is assigned to one of the categories specified by $\mathcal{T}$. These categories can be dynamically defined by natural-language descriptions and may include concepts unseen during training.

In this work, our OWL-Net aims to improve open-vocabulary segmentation under low-light conditions. To this end, OWL-Net additionally exploits the corresponding RAW data $I_{raw} \in \mathbb{R}^{4\times H/2\times W/2}$ during training. RAW data preserve finer intensity variations and structural details that are often weakened in low-light RGB images, and therefore provide useful guidance for learning robust visual representations. Importantly, RAW data are only used during training, while inference relies solely on the low-light RGB image. In the following parts, we first describe the two core components of OWL-Net in detail, and then introduce the overall pipeline of our method.

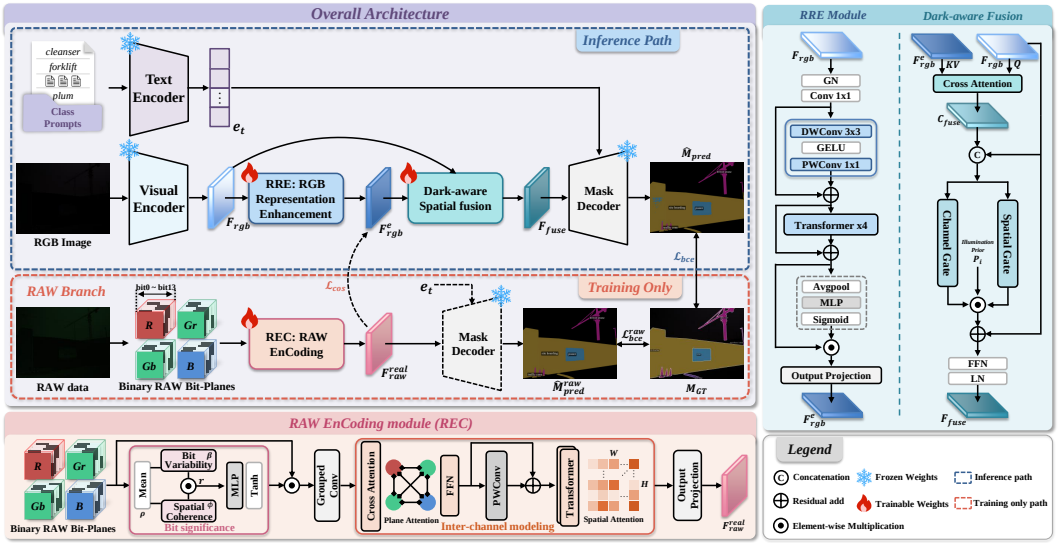



Figure 2: Overall framework of our OWL-Net. Notably, the RAW branch is only utilized during training to facilitate the visual feature learning under low-light conditions. Once the training is done, OWL-Net only relies on the RGB image for inference.

### 3.2 RAW ENCODING MODULE

RAW data retain finer intensity variations and structural details than rendered RGB images under low-light conditions, making them valuable for improving visual representation learning. However, directly incorporating RAW data into deep neural networks is non-trivial. In contrast to conventional 8-bit RGB images, RAW measurements are typically represented with a considerably higher bit depth, resulting in a much larger numerical range. Directly compressing RAW values into a standard value range may lose the subtle differences carried by different bit positions, thereby weakening the fine-grained information carried by the higher bit depth. To better preserve such details, we introduce a RAW EnCoding (REC) module, which explicitly processes the high-magnitude RAW data at the bit-plane level and thus preserves high-bit-depth information during feature extraction.

Given a 14-bit RAW image $I_{raw}$ captured with a four-channel RGGB Bayer pattern, we decompose each Bayer channel into 14 binary bit planes, denoted collectively by $I^c_{raw} = \{I^c_{raw,k}\}_{k=0}^{13}$, where $c \in \{1, 2, 3, 4\}$ indexes the Bayer channel and $k$ indexes the bit position. The REC module estimates the significance of each bit according to two complementary properties: bit variability and spatial coherence. Specifically, for each bit plane $I^c_{raw,k}$, we first compute its occupancy $\rho$ and define the bit variability score $\beta$ as:

$$\rho = \text{avg}(I^c_{raw,k}), \qquad \beta = 4\rho(1-\rho). \tag{1}$$

Here, $\rho$ denotes the proportion of activated bits in the corresponding bit plane. The bit variability score $\beta$ is derived from the Bernoulli variance and normalized to $[0, 1]$. It assigns lower scores to bit planes strongly dominated by either 0 or 1, as such planes provide limited fine-grained structural information. However, a high bit variability score alone does not necessarily indicate useful structure, since random noise with a balanced distribution of zeros and ones can also yield a high value of $\beta$. We therefore further measure the spatial coherence of each bit plane. To this end, we calculate the spatial variation of each bit plane using a direction-averaged anisotropic variation:

$$d = \tfrac{1}{2}\big(\text{avg}\,|\Delta_x I^c_{raw,k}| + \text{avg}\,|\Delta_y I^c_{raw,k}|\big), \tag{2}$$

where $\Delta_x$ and $\Delta_y$ denote horizontal and vertical first-order differences, respectively. For binary bit planes, $d$ measures the average disagreement rate between neighboring bits. To assess spatial coherence relative to random fluctuations, we normalize $d$ by the expected disagreement $2\rho(1-\rho)$ at the same occupancy. Therefore, the spatial coherence score is defined as (Gallego et al., 2019):

$$\phi = \text{clip}_{[0,1]}\left(1 - \frac{d}{2\rho(1-\rho)+\epsilon}\right), \tag{3}$$

where $\epsilon$ is a small positive constant introduced for numerical stability. Random fluctuation noise tends to produce a low coherence score, while spatially continuous structures receive a higher score. We then combine the bit variability score $\beta$ and the spatial coherence score $\phi$ to obtain the significance score of each bit plane as $r = \beta \cdot \phi$. For the $c$-th Bayer channel, the significance scores of its 14 bit planes are collected as $r^c = [r_0^c, \ldots, r_{13}^c]$, which is further mapped to a bit significance vector via $v^c = \mathrm{Tanh}(\mathrm{MLP}(r^c))$. The generated $v^c$ is used to reweight the corresponding 14 bit planes and produce $\tilde{I}_{raw}$, which emphasizes the informative and spatially coherent bit-level responses while assigning lower weights to nearly constant or noise-dominated ones.

Afterwards, $\tilde{I}_{raw}$ is first projected by a $1 \times 1$ grouped convolution to independently embed each Bayer channel. Then, cross-attention is applied across the four channel-wise representations to exploit inter-channel interactions. Subsequently, the generated features from the four Bayer channels are concatenated along the channel dimension and further refined by a pointwise convolution. Finally, a Transformer encoder is employed to capture the long-range dependencies across spatial locations, followed by a projection layer to produce the encoded RAW representation $F_{raw}^{real}$, which is regarded as the final output of REC module.

### 3.3 RGB Representation Enhancement Module

Although RAW data provide richer structural and intensity information under low-light conditions, requiring RAW input during inference will limit the flexible deployment of the model. Therefore, we introduce an RGB Representation Enhancement (RRE) module, which converts the degraded RGB features into the RAW-aware representation space to recover the overwhelmed details.

Given the RGB feature $F_{rgb}$ extracted by the visual encoder, RRE progressively enhances it through two residual stages in a local-to-global manner. We first employ a lightweight local compensation block consisting of a $3 \times 3$ depthwise convolution, a GELU activation, and a $1 \times 1$ pointwise convolution. The predicted residual is added back to the input feature, allowing the module to compensate for local structural cues that are severely weakened under extreme low-light conditions. This process can be expressed as:

$$F_{local} = F_{rgb} + \mathcal{H}_{local}(F_{rgb}), \tag{4}$$

where $\mathcal{H}_{local}$ denotes the local compensation block. Based on the locally enhanced representation $F_{local}$, we further employ four Transformer blocks to capture long-range dependencies and recover the global contextual information. Similarly, the global compensation block is also formulated in a residual manner:

$$F_{global} = F_{local} + \mathcal{H}_{global}(F_{local}), \tag{5}$$

where $\mathcal{H}_{global}$ denotes the stacked Transformer blocks. In this way, RRE first compensates for locally degraded structures and subsequently models their global dependencies, progressively lifting the degraded RGB representation toward the RAW-aware feature space.

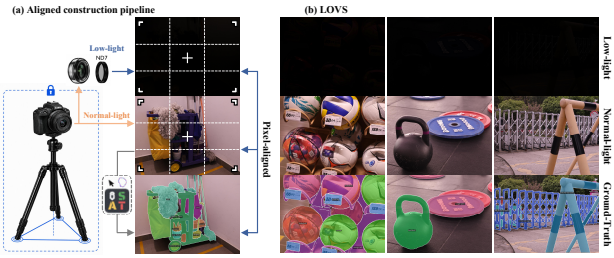
Finally, we introduce a lightweight feature refinement block to adaptively recalibrate the enhanced representation via the channel-wise attention, which emphasizes informative channels while suppressing less relevant ones. An output projection is then applied to obtain the final enhanced RGB representation $F_{rgb}^{e}$. During training, $F_{rgb}^{e}$ is supervised by the real RAW representation $F_{raw}^{real}$ produced by the REC module through a cosine alignment loss:

$$\mathcal{L}_{cos} = 1 - \frac{\left\langle F_{rgb}^{e}, \mathrm{sg.}(F_{raw}^{real}) \right\rangle}{\left\| F_{rgb}^{e} \right\|_2 \left\| \mathrm{sg.}(F_{raw}^{real}) \right\|_2}, \tag{6}$$

where $\mathrm{sg.}(\cdot)$ denotes the stop-gradient operation, which prevents the gradients from flowing back into the RAW encoder and avoids the undesired distortion of the RAW features. Through this progressive learning process, the RRE module gradually promotes the degraded RGB features toward the representation encoded from real RAW data, recovering visual cues weakened under low-light conditions. Consequently, OWL-Net learns to produce detail-enhanced representations directly from RGB features and therefore does not require RAW input during inference.

### 3.4 Overall pipeline

The overall pipeline of OWL-Net is illustrated in Fig. 2. During training, the RAW image is first fed into the REC module to obtain the encoded RAW representation $F_{raw}^{real}$, while the low-light RGB




| Dataset | #Cls | #Img | mRL ↓ | mHSV-V ↓ | mCS ↓ | Dataset modalities | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | Normal | Dark | RAW |
| *Open-vocabulary datasets* | | | | | | | | |
| VOC-20 | 20 | 1,449 | 0.244 | 0.501 | 0.98 | ✓ | ✗ | ✗ |
| PC-59 | 59 | 5,105 | 0.247 | 0.501 | 0.95 | ✓ | ✗ | ✗ |
| ADE-150 | 150 | 2,000 | 0.262 | 0.532 | 0.81 | ✓ | ✗ | ✗ |
| PC-459 | 459 | 5,105 | 0.247 | 0.501 | 0.84 | ✓ | ✗ | ✗ |
| Cityscapes | 19 | 1,525 | 0.111 | 0.316 | 0.85 | ✓ | ✗ | ✗ |
| *Low-light perception datasets* | | | | | | | | |
| LIS | 8 | 669 | 0.004 | 0.035 | 1.00 | ✓ | ✓ | ✓ |
| ExDark | 12 | 2,563 | 0.048 | 0.175 | 0.98 | ✗ | ✓ | ✗ |
| ACDC-Night | 19 | 500 | 0.054 | 0.235 | 0.92 | ✓ | ✓ | ✗ |
| Dark Zurich | 19 | 151 | 0.055 | 0.235 | 0.92 | ✓ | ✓ | ✗ |
| NightCity | 19 | 1,299 | 0.066 | 0.230 | 0.92 | ✗ | ✓ | ✗ |
| LOVS (Ours) | 220 | 1,000 | 0.001 | 0.019 | 0.78 | ✓ | ✓ | ✓ |

Figure 3: Overview of the LOVS dataset. **Left:** (a) The illustration of paired normal-/low-light image collection and annotation pipeline. (b) Representative samples from LOVS. **Right:** Comparison with representative open-vocabulary and low-light perception datasets.

image is processed by the visual encoder to extract the RGB feature $F_{rgb}$. The extracted $F_{rgb}$ is then passed through the RRE module to produce an enhanced representation $F^e_{rgb}$, which captures richer structural and semantic cues that are difficult to preserve in extremely low-light RGB images.

Afterwards, $F^e_{rgb}$ is integrated with the original RGB feature $F_{rgb}$ through a dark-aware spatial fusion module to produce the aggregated feature $F_{fuse}$. Meanwhile, given the class prompts, the text encoder extracts the corresponding textual embeddings $e_t$. Conditioned on $e_t$, the mask decoder takes $F_{fuse}$ as visual input and predicts the final segmentation mask $\hat{M}_{pred}$. To ensure that the RAW representation itself preserves task-relevant semantic information, $F^{real}_{raw}$ is also fed into the mask decoder during training and produces the prediction $\hat{M}^{raw}_{pred}$. Accordingly, based on the ground-truth segmentation mask $M_{gt}$, we compute the binary cross-entropy losses $\mathcal{L}_{bce}$ and $\mathcal{L}^{raw}_{bce}$ for $\hat{M}_{pred}$ and $\hat{M}^{raw}_{pred}$, respectively. The overall training objective is the combination of $\mathcal{L}_{bce}$, $\mathcal{L}^{raw}_{bce}$, and $\mathcal{L}_{cos}$.

Once training is completed, the RAW input and the RAW branch are entirely removed. At inference, OWL-Net only relies on the low-light RGB image. The RRE module lifts the illumination-degraded RGB feature toward the RAW-aware representation space to recover the overwhelmed details. Thus, the mask decoder receives a more informative visual representation for matching with textual embeddings, enabling more reliable visual-text correspondence and thus improving open-vocabulary segmentation under low-light conditions. Finally, $\hat{M}_{pred}$ is utilized as the model prediction.

# 4 LOVS Dataset

## 4.1 Construction Pipeline

To construct the LOVS dataset, we first acquire geometrically aligned normal-/low-light image pairs, and then we conduct annotations on the normal-light images.

**Data Collection.** We collect paired normal-light and low-light images from diverse real-world scenes. As illustrated in Fig. 3, for each scene, we first capture a reference image under normal illumination with the Canon EOS R50 camera mounted on a fixed tripod. Then, without changing the camera position and viewpoint, we recapture the same scene under low-light conditions by introducing a 7-stop neutral-density (ND) filter, which substantially attenuates the incident light by 128×. During the paired acquisition, the camera pose and major imaging configurations are kept unchanged to minimize geometric misalignment. In this manner, each low-light image is paired with a normal-light counterpart, allowing the more reliable annotation in the following steps. In total, we collect 1,000 paired normal-light and low-light images, and the corresponding RAW data are simultaneously recorded during low-light image acquisition.

**Data Annotation.** During annotation, we first define the category taxonomy, which contains 220 semantic categories spanning 10 different domains: Living, Food, Cleaning, Office, Retail, Sports, Mobility, Construction, Utilities, and Safety. Then, for each retained sample, annotation is performed on the corresponding normal-light image. Annotators first identify the categories present in the image according to the predefined taxonomy. The category names are then provided to the ISAT-SAM annotation tool (Ji & Zhang, 2024) to generate initial segmentation masks. Afterwards, we manually inspect all the automatically generated masks and refine the inaccurate mask regions, producing the final pixel-level annotations. Since each normal-light image is geometrically aligned

Table 1: The comparisons of our OWL-Net with different methods. Notably, all the models are trained on the synthetic low-light COCO-Stuff and then evaluated on the different datasets.

| Model | Visual Backbone | Text Encoder | ADE-150 | | ADE-847 | VOC-20 | PC-59 | PC-459 | LOVS | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | mIoU↑ | PQ↑ | mIoU↑ | mIoU↑ | mIoU↑ | mIoU↑ | mIoU↑ | hIoU↑ |
| OVSeg (Liang et al., 2023) | ViT-B/16 | CLIP | 8.6 | 3.6 | 1.8 | 51.6 | 23.4 | 3.3 | 5.5 | 7.0 |
| FC-CLIP (Yu et al., 2023) | ResNet-101 | CLIP | 10.4 | 6.8 | 2.5 | 57.4 | 21.1 | 3.5 | 5.0 | 6.6 |
| OpenSeeD (Zhang et al., 2023) | Swin-T | UniCL | 9.2 | 5.7 | 2.0 | 51.9 | 27.5 | 4.6 | 7.2 | 7.6 |
| X-Decoder (Zou et al., 2023) | FocalNet-T | UniCL | 13.5 | 6.3 | 2.3 | 56.6 | 30.8 | 4.8 | 7.4 | 8.2 |
| SAN (Xu et al., 2023) | ViT-B/16 | CLIP | 12.7 | – | 4.6 | 64.8 | 30.7 | 6.4 | 12.2 | 12.5 |
| MAFT (Jiao et al., 2023) | ViT-B/16 | CLIP | 16.2 | 9.9 | 3.4 | 69.0 | 30.8 | 5.8 | 9.5 | 10.8 |
| MAFT+ (Jiao et al., 2024) | ConvNeXt-B | CLIP | 20.8 | 12.1 | 5.1 | 65.6 | 33.7 | 6.4 | 11.0 | 11.6 |
| SED (Xie et al., 2024) | ConvNeXt-B | CLIP | 17.2 | – | 4.1 | 65.2 | 32.5 | 6.2 | 14.5 | 14.3 |
| MaskAdapter (Li et al., 2025a) | ConvNeXt-B | CLIP | 22.0 | 12.7 | 6.2 | 73.1 | 37.1 | 8.4 | 12.7 | 12.8 |
| CAT-Seg (Cho et al., 2024) | ViT-B/16 | CLIP | 18.0 | – | 5.9 | 71.6 | 35.5 | 8.3 | 14.9 | 15.6 |
| OWL-Net (CAT-Seg) | ViT-B/16 | CLIP | 21.6 | – | 6.6 | 72.4 | 38.4 | 8.8 | 16.7 | 16.9 |
| OVSeg (Liang et al., 2023) | ViT-L/14 | CLIP | 12.0 | 6.4 | 2.9 | 72.7 | 33.4 | 5.2 | 12.4 | 15.3 |
| FC-CLIP (Yu et al., 2023) | ConvNeXt-L | CLIP | 14.8 | 9.1 | 4.2 | 77.5 | 32.4 | 6.3 | 16.6 | 19.2 |
| OpenSeeD (Zhang et al., 2023) | Swin-L | UniCL | 16.8 | 9.5 | 4.2 | 78.2 | 45.7 | 8.9 | 12.1 | 14.2 |
| X-Decoder (Zou et al., 2023) | FocalNet-L | UniCL | 20.9 | 10.3 | 4.5 | 86.8 | 47.1 | 8.2 | 13.4 | 16.1 |
| SAN (Xu et al., 2023) | ViT-L/14 | CLIP | 18.9 | – | 7.2 | 78.4 | 39.0 | 9.6 | 20.7 | 20.3 |
| MAFT (Jiao et al., 2023) | ViT-L/14 | CLIP | 20.1 | 10.8 | 5.1 | 80.4 | 35.0 | 7.5 | 17.7 | 18.3 |
| MAFT+ (Jiao et al., 2024) | ConvNeXt-L | CLIP | 23.5 | 13.7 | 6.2 | 72.1 | 37.5 | 7.1 | 16.8 | 17.9 |
| SED (Xie et al., 2024) | ConvNeXt-L | CLIP | 23.2 | – | 7.7 | 76.6 | 38.9 | 9.6 | 26.6 | 23.2 |
| MaskAdapter (Li et al., 2025a) | ConvNeXt-L | CLIP | 25.0 | 15.2 | 7.3 | 77.4 | 41.5 | 10.2 | 22.1 | 22.5 |
| OVRCOAT (Kormushev et al., 2026) | ConvNeXt-L | CLIP | 25.3 | 16.2 | 7.9 | 75.3 | 35.0 | 8.6 | 22.6 | 21.4 |
| CAT-Seg (Cho et al., 2024) | ViT-L/14 | CLIP | 26.4 | – | 9.7 | 85.0 | 46.9 | 13.1 | 28.0 | 27.0 |
| OpenWorldSAM (Xiao et al., 2026) | SAM2 Hiera-L | BEiT-3 | 34.7 | 15.1 | 18.1 | 91.3 | 51.8 | 22.8 | 48.1 | 46.0 |
| OWL-Net (OVRCOAT) | ConvNeXt-L | CLIP | 28.0 | 17.8 | 8.2 | 75.3 | 39.3 | 9.3 | 27.5 | 26.5 |
| OWL-Net (CAT-Seg) | ViT-L/14 | CLIP | 28.0 | – | 9.8 | 87.0 | 49.3 | 13.5 | 29.3 | 28.1 |
| OWL-Net (OpenWorldSAM) | SAM2 Hiera-L | BEiT-3 | **37.3** | **16.5** | **19.6** | **92.3** | **53.6** | **25.4** | **49.2** | **51.3** |

with its low-light counterpart during data collection, the annotated masks are directly transferred to the corresponding low-light images and used as ground-truth annotations.

### 4.2 Dataset Statistics

In total, LOVS contains 1,000 paired normal-light and low-light images covering diverse indoor and outdoor scenes. After annotation, the dataset includes 6,949 object-level masks across 220 semantic categories, providing a comprehensive benchmark for evaluating open-vocabulary segmentation under challenging illumination conditions. As summarized in the right part of Fig. 3, we compare LOVS with representative open-vocabulary segmentation datasets and low-light perception benchmarks. Notably, since our LOVS is only utilized for evaluation, we only select the test sets of the existing datasets for comparison. As shown, LOVS provides relatively broader category coverage while maintaining a considerable evaluation scale. Fig. 3 also reports the mean relative luminance (mRL) (Commission et al., 1999; Caldwell et al., 2008) and mean HSV value (mHSV-V) (Tan et al., 2021), two commonly used metrics for measuring image brightness. LOVS has the lowest values on both metrics, reflecting the challenging low-light conditions considered in our benchmark. In addition, we compute mean category similarity (mCS) (Liu et al., 2025) to measure the distribution similarity between each evaluation dataset and the COCO-Stuff (Lin et al., 2014) training set. A lower mCS indicates lower similarity to COCO-Stuff and thus a larger distribution gap from the training categories. As shown in Fig. 3, LOVS receives the lowest mCS among the compared datasets, suggesting that its semantic content is less similar to the training set and therefore provides a more challenging evaluation for open-vocabulary generalization.

## 5 Experiments

### 5.1 Experimental Setup

**Training data.** Due to the lack of low-light images with open-vocabulary segmentation annotations, all models in our experiments are trained on synthetic low-light data. Specifically, we follow the standard open-vocabulary segmentation and utilize COCO-Stuff as the training data. According to

Table 2: The effectiveness analysis of different components within the REC module.

| Variant | LOVS | |
|---|---|---|
| | mIoU | hIoU |
| OpenWorldSAM | 48.1 | 46.0 |
| Direct Compression | 47.6 | 48.9 |
| w/ Bit Significance | 48.4 | 49.7 |
| w/ Inter-channel Modeling | 48.9 | 50.2 |
| Full REC | **49.2** | **51.3** |

Table 3: The effectiveness analysis of different components within the RRE module.

| Variant | LOVS | |
|---|---|---|
| | mIoU | hIoU |
| OpenWorldSAM | 48.1 | 46.0 |
| Plain Conv. | 48.1 | 48.8 |
| w/ Local block | 48.8 | 50.6 |
| w/ Global block | 48.5 | 49.9 |
| Params-matched Conv. | 48.8 | 50.8 |
| Full RRE | **49.2** | **51.3** |
| Including Real RAW Input | **49.5** | **51.9** |

the strategy in (Wei et al., 2020), we employ the Physics-based Noise Formation Model to synthesize low-light RGB images and their corresponding RAW data based on COCO-Stuff images. In total, all 118,287 COCO-Stuff images are converted into low-light counterparts and used for training.

**Evaluation datasets and metrics.** To provide a comprehensive evaluation, we conduct experiments on both synthetic and real low-light open-vocabulary segmentation datasets. On the one hand, to evaluate each model in the real low-light scenarios, we conduct experiments on our newly collected LOVS dataset. On the other hand, we also perform evaluation on five widely used open-vocabulary segmentation datasets, including ADE20K-150/847 (Zhou et al., 2017; 2019), PASCAL VOC-20 (Everingham et al., 2010), and PASCAL Context-59/459 (Mottaghi et al., 2014). Similar to the strategy used for training data, we synthesize the low-light counterparts for these datasets using the Physics-based Noise Formation Model (Wei et al., 2020). During the evaluation on our LOVS dataset, we report mean Intersection over Union (mIoU) and harmonic mean Intersection over Union (hIoU) to assess the segmentation performance. For the five synthetic datasets, we follow the standard evaluation protocols of the corresponding datasets and report their commonly used metrics.

**Implementation details.** We implement OWL-Net on different representative open-vocabulary segmentation frameworks, including the CAT-Seg, OVRCOAT and OpenWorldSAM. For each framework, we keep the original model configurations unchanged, and only insert the proposed components of OWL-Net. This setting allows us to evaluate whether OWL-Net can consistently improve different open-vocabulary segmentation frameworks in a plug-and-play manner. Notably, for fair comparisons, all the models are trained according to their original hyper-parameters. More implementation details are provided in Appendix E.

### 5.2 Comparisons with state-of-the-art Methods

In this section, we compare representative open-vocabulary segmentation methods under low-light conditions. The results are summarized in Table 1, which is organized into two groups along the rows. The upper part reports methods built on relatively smaller backbones, while the lower part includes methods using larger backbones. It is important to emphasize that, for a fair comparison, all the compared methods in Table 1 are re-trained on synthetic low-light COCO-Stuff, which is the same as our OWL-Net.

As shown in Table 1, OWL-Net achieves the best performance across different datasets. In particular, on LOVS, OWL-Net outperforms the second-best method by 1.8% and 1.3% in terms of mIoU and hIoU when equipped with base-size visual encoders. With larger backbones, OWL-Net also achieves clear improvements of 1.1% and 5.3% on the two metrics. Moreover, when integrated with CAT-Seg, OVRCOAT, and OpenWorldSAM, OWL-Net consistently improves the performance of different open-vocabulary segmentation frameworks, which demonstrates the plug-and-play capability of our method across different architectures. Fig. 4(a) illustrates the segmentation map generated by different methods, in which OWL-Net provides more accurate segmentation results under low-light conditions. Notably, OWL-Net only uses RAW data to facilitate feature learning during training. Once training is completed, the RAW branch is removed and inference is performed using RGB input only. This design avoids the additional transmission and storage overhead of RAW data during deployment, leading to a more efficient and flexible inference pipeline.

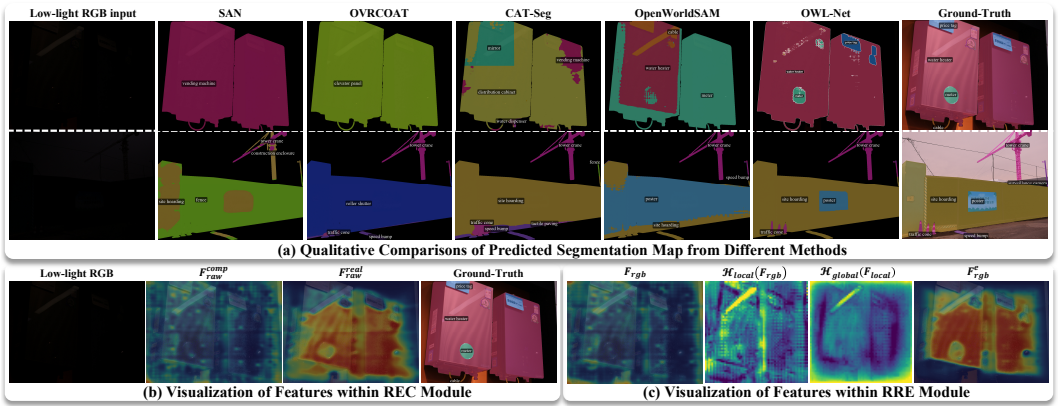



Figure 4: Qualitative results and feature visualizations. **(a)** Segmentation results of different open-vocabulary methods under low-light conditions. **(b)** Directly compressed and REC-encoded RAW features. **(c)** Visualizations of intermediate features within the RRE module.

### 5.3 Ablation Study

We conduct ablation studies to analyze the effectiveness of each component within OWL-Net. All the experiments are performed on the LOVS dataset using OpenWorldSAM as the basic framework.

In Table 2, we evaluate different strategies for encoding RAW data. As shown, directly compressing RAW values (Feijoo et al., 2025) yields the weakest performance among the RAW-based variants, providing only a marginal improvement over the plain OpenWorldSAM baseline. When bit significance estimation and inter-channel modeling are introduced, respectively, the performance is consistently improved. The full REC module achieves the best result, reaching 49.2 mIoU and 51.3 hIoU on LOVS, demonstrating the effectiveness of two core components within our REC module. Meanwhile, Fig. 4**(b)** visualizes the RAW features encoded by different approaches. Compared with the directly compressed feature $F_{raw}^{comp}$, the REC-encoded feature $F_{raw}^{real}$ produces more concentrated responses on semantically relevant object regions and better preserves their structural details.

In Table 3, we investigate the benefit of each component within the RRE module. Compared with the plain baseline, both the local compensation block and the global compensation block obtain consistent performance improvements. To further exclude the possibility that the performance gains mainly arise from the increased parameters, we construct a baseline by stacking convolutional layers with a comparable number of parameters to the full RRE module, which still performs worse than our model. These results demonstrate that our RRE benefits from the proposed local-to-global compensation mechanism, which effectively recovers visual details weakened in low-light RGB features. Moreover, Fig. 4**(c)** visualizes the intermediate features within RRE module. The local compensation block enhances fine-grained structures and object boundaries, while the global compensation block captures broader spatial dependencies. After progressive enhancement, the final representation $F_{rgb}^{e}$ exhibits more concentrated responses on semantically relevant object regions, which is qualitatively close to the corresponding real RAW feature in Fig. 4**(b)**. Furthermore, as listed in Table 3, including real RAW data during inference only brings marginal improvements over our method. This demonstrates that RRE can effectively learn informative representations from degraded RGB features, which closely approximate those provided by real RAW data.

## 6 Conclusion

In this work, we study open-vocabulary segmentation under challenging low-light conditions and propose OWL-Net, which leverages RAW data during training while requiring only RGB input during inference. OWL-Net introduces the RAW EnCoding (REC) module to preserve informative bit-level details from high-bit-depth RAW data, and the RGB Representation Enhancement (RRE) module to strengthen degraded RGB representations with RAW guidance. We also construct LOVS, a real-world low-light open-vocabulary segmentation dataset with diverse categories and pixel-level annotations. Extensive experiments on both synthetic benchmarks and LOVS demonstrate the effectiveness and generality of OWL-Net across different open-vocabulary segmentation frameworks.

## AI USE STATEMENT

Generative AI tools were used only to aid and polish the writing of this paper. They were not used for research methodology, experiment design, data generation or analysis, method implementation, mathematical formulation, or result interpretation. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final manuscript.

## ETHICS STATEMENT

This work does not involve human-subject research or the collection of personally identifiable information. The LOVS dataset consists of real-world scene images collected for low-light open-vocabulary segmentation, and the data collection was designed to avoid capturing identifiable individuals or other sensitive personal information. We will further inspect the dataset before release and remove any samples that may raise privacy concerns. The dataset will be released for research purposes under appropriate usage terms. We have read and followed the ICLR Code of Ethics and take responsibility for the ethical conduct of this work.

## REFERENCES


Jong-Hyeon Baek, Jiwon Oh, and Yeong Jun Koh. Evolve: Event-guided deformable feature transfer and dual-memory refinement for low-light video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11273–11282. IEEE, 2025.

Qi Bi, Jingjun Yi, Huimin Huang, Hao Zheng, Haolan Zhan, Yawen Huang, Yuexiang Li, Xian Wu, and Yefeng Zheng. Nightadapter: Learning a frequency adapter for generalizable night-time scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23838–23849. IEEE, 2025.

Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

Ben Caldwell, Michael Cooper, Loretta Guarino Reid, Gregg Vanderheiden, Wendy Chisholm, John Slatin, and Jason White. Web content accessibility guidelines (wcag) 2.0. *WWW Consortium (W3C)*, 290(1-34):5–12, 2008.

Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4113–4123. IEEE, 2024.

International Electrotechnical Commission et al. Multimedia systems and equipment-color measurement and management-part 2-1. *Color management-Default RGB color space-sRGB*, 1999.

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. IEEE, 2025.

Gourav Datta, Zeyu Liu, Zihan Yin, Linyu Sun, Akhilesh R Jaiswal, and Peter A Beerel. Enabling ispless low-power computer vision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2429–2438. IEEE, 2023.

Tian Ding, Hongtao Yang, Liangtao Shi, Jun Li, Xiantao Hu, Jian Yang, and Ying Tai. Adaptive depth lightweight rgb-t tracking with holistic token routing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20942–20952, 2026.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2010.

Huijie Fan, Yu Qiao, Yihao Zhen, Tinghui Zhao, Baojie Fan, and Qiang Wang. All-day multi-camera multi-target tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16892–16901. IEEE, 2025.

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*, pp. 10879–10889. IEEE, 2025.

Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12272–12281. IEEE, 2019.

Jiasheng Guo, Xin Gao, Yuxiang Yan, Guanghao Li, and Jian Pu. Dark-isp: Enhancing raw image processing for low-light object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9583–9593. IEEE, 2025.

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.

Khurram Azeem Hashmi, Karthik Palyakere Suresh, Didier Stricker, and Muhammad Zeshan Afzal. Torchadapt: Towards light-agnostic real-time visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5645–5656. IEEE, 2025.

Chunming He, Chengyu Fang, Yulun Zhang, Longxiang Tang, Jinfa Huang, Kai Li, Xiu Li, Sina Farsiu, et al. Reti-diff: Illumination degradation image restoration with retinex-based latent diffusion model. In *International Conference on Learning Representations*, volume 2025, pp. 43332–43352, 2025.

Shuwei Ji and Hongyuan Zhang. ISAT with Segment Anything: An Interactive Semi-Automatic Annotation Tool, 2024. URL `https://github.com/yatengLG/ISAT_with_segment_anything`. Updated on 2025-02-07.

Hai Jiang, Binhao Guan, Zhen Liu, Xiaohong Liu, Jian Yu, Zheng Liu, Songchen Han, and Shuaicheng Liu. Learning to see in the extremely dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7676–7685. IEEE, 2025.

Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021.

Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 36:35631–35653, 2023.

Siyu Jiao, Hongguang Zhu, Jiannan Huang, Yao Zhao, Yunchao Wei, and Humphrey Shi. Collaborative vision-text representation optimizing for open-vocabulary segmentation. In *European Conference on Computer Vision*, pp. 399–416. Springer, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3992–4003. IEEE, 2023.

Nikolay Kormushev, Josip Šarić, and Matej Kristan. Mitigating objectness bias and region-to-text misalignment for open-vocabulary panoptic segmentation. *arXiv preprint arXiv:2603.21386*, 2026.

Minhyeok Lee, Suhwan Cho, Jungho Lee, Sunghun Yang, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Effective sam combination for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26081–26090. IEEE, 2025.

Hebei Li, Jin Wang, Jiahui Yuan, Yue Li, Wenming Weng, Yansong Peng, Yueyi Zhang, Zhiwei Xiong, and Xiaoyan Sun. Event-assisted low-light video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3250–3259. IEEE, 2024.

Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, and Xinggang Wang. Mask-adapter: The devil is in the masks for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14998–15008. IEEE, 2025a.

Zhong-Yu Li, Xin Jin, Bo-Yuan Sun, Chun-Le Guo, and Ming-Ming Cheng. Towards raw object detection in diverse conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8859–8868. IEEE, 2025b.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7070. IEEE, 2023.

Tsung-Yi Lin, Michael Maire, Serge 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. Springer, 2014.

Yong Liu, Song-Li Wu, Sule Bai, Jiahao Wang, Yitong Wang, and Yansong Tang. Stepping out of similar semantic space for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22664–22675. IEEE, 2025.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 891–898, 2014.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Xin Tan, Ke Xu, Ying Cao, Yiheng Zhang, Lizhuang Ma, and Rynson WH Lau. Night-time scene parsing with a large real dataset. *IEEE Transactions on Image Processing*, 30:9085–9098, 2021.

Xinzhe Wang, Kang Ma, Qiankun Liu, Yunhao Zou, and Ying Fu. Multi-object tracking in the dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 382–392. IEEE, 2024.

Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2755–2764. IEEE, 2020.

Zhixiang Wei, Lin Chen, Tao Tu, Pengyang Ling, Huaian Chen, and Yi Jin. Disentangle then parse: Night-time semantic segmentation with illumination disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21536–21546. IEEE, 2023.

Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, et al. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):5092–5113, 2024.

Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15764–15773. IEEE, 2021.

Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8248–8257. IEEE, 2019.

Shiting Ginny Xiao, Rishabh Kabra, Yuhang Li, Donghyun Lee, Joao Carreira, and Priyadarshini Panda. Openworldsam: Extending sam2 for universal image segmentation with language prompts. *Advances in Neural Information Processing Systems*, 38:35507–35532, 2026.

Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3426–3436. IEEE, 2024.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: Side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15546–15561, 2023.

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. IEEE, 2025.

Siyuan Yao, Rui Zhu, Ziqi Wang, Wenqi Ren, Yanyang Yan, and Xiaochun Cao. Umdatrack: Unified multi-domain adaptive tracking under adverse weather conditions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6466–6475. IEEE, 2025.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023.

Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. In *European Conference on Computer Vision*, pp. 419–437. Springer, 2024.

Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1020–1031. IEEE, 2023.

Pengzhi Zhong, Xiaoyu Guo, Defeng Huang, Xiaojun Peng, Yian Li, Qijun Zhao, and Shuiwang Li. Low-light object tracking: A benchmark. *IEEE Transactions on Intelligent Vehicles*, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5122–5130. IEEE, 2017.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 8954–8975, 2024.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15116–15127. IEEE, 2023.

# APPENDIX

## A ADDITIONAL VISUALIZATION ANALYSIS

### A.1 MOTIVATION FOR RAW-ASSISTED REPRESENTATION LEARNING

To illustrate the benefit of using RAW data to facilitate the perception under low-light conditions, we visualize the binary bit planes of low-light RGB and RAW data from the same scenes. As shown in Fig. 5, the visible structural details in low-light RGB images are mainly concentrated in bit plane 0, 1 and 2, while bit plane 3 contains limited details and the higher bit planes exhibit few visible responses. In comparison, RAW data preserve clearer object structures across multiple bit positions, particularly in bit planes 2 to 8. Meanwhile, bit planes 9 to 11 still retain discernible visual information. These observations suggest that RAW measurements retain richer information that is weakened in rendered low-light RGB images, which motivates us to use such information to guide RGB representation learning. Meanwhile, the noise-like responses in the lower RAW bit planes further highlight the importance of identifying informative bit planes rather than treating all bit positions equally.

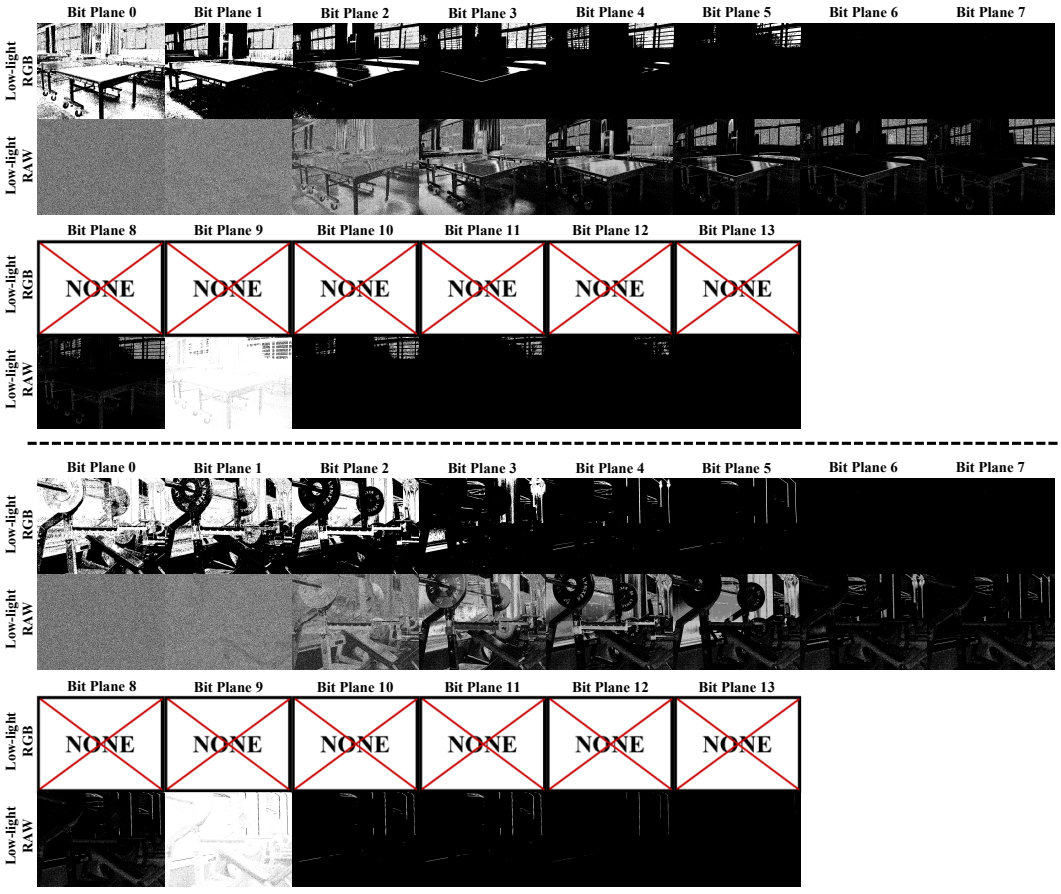



Figure 5: Bit-plane visualization of low-light RGB and RAW data. As shown, two representative samples from LOVS are selected for visualization, with bit-plane indices starting from the least significant bit (LSB). Compared with low-light RGB images, RAW data preserve visible structural details across a wider range of bit planes, providing complementary information for low-light representation learning.

### A.2 Visualization of Correspondence Score

To further investigate how the RRE module improves visual-text correspondence under low-light conditions, we visualize the pixel-wise similarity scores between visual features and textual embeddings involved in cross-attention. As shown in Fig. 6, the features $F_{rgb}$ extracted from the low-light RGB image produce high similarity scores with the textual embedding in both foreground and background regions, making it difficult to establish reliable correspondences between image regions and textual concepts. In contrast, the enhanced feature $F^e_{rgb}$ generated by RRE yields more concentrated responses on the target objects while suppressing irrelevant background responses. The change of correspondence scores further highlights the improved localization of semantically relevant regions. The experimental results suggest that the visual feature $F^e_{rgb}$ produced by the RRE module contains more discriminative information, enabling more reliable correspondences between image regions and textual concepts to be established. Therefore, OWL-Net achieves better open-vocabulary segmentation performance under low-light conditions.

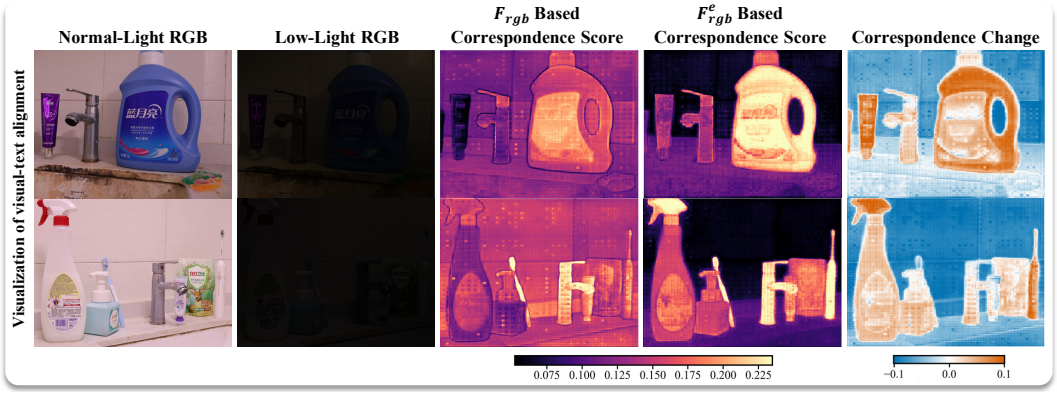



Figure 6: Visualization of visual–text correspondence under low-light conditions. In the cross-attention operation, pixel-wise correspondence scores are computed between textual embeddings and the original RGB features $F_{rgb}$ or the RRE-enhanced features $F^e_{rgb}$. The enhanced features produce more concentrated responses on target objects while reducing irrelevant background responses, facilitating more reliable visual–text interactions through cross-attention.

### A.3 The Analysis of Bit Significance within REC module

To further analyze the bit significance estimation in the REC module, we provide visualizations from two aspects. First, we examine the structural information preserved across different bit planes and Bayer channels. As illustrated in Fig. 7, different bit planes exhibit noticeable differences in object structures, boundaries, and fine-grained details. Moreover, even at the same bit position, the R, Gr, Gb, and B channels show different structural responses. These observations motivate us to estimate bit significance independently across both Bayer channels and bit positions, allowing each bit plane in each channel to be weighted individually.

To further examine whether the proposed significance estimation mechanism can identify informative bit planes, we visualize the bit variability score $\beta$, spatial coherence score $\phi$, and their combined significance score $r = \beta \cdot \phi$ for each RAW bit plane in Fig. 7. As shown, $\beta$ assigns higher scores to bit planes with a more balanced distribution of zeros and ones, while those dominated by either zeros or ones receive lower scores. Meanwhile, $\phi$ assigns lower scores to bit planes dominated by random fluctuations and higher scores to those containing spatially coherent structures. By combining these two properties, the resulting significance score $r$ exhibits the highest responses in bit planes 2 to 5, where clearer object structures and fine-grained details are preserved. This observation is consistent with the bit-plane visualization in Sec. A.1, where RAW bit planes 2 to 5 retain richer structural information and therefore play more important roles during feature extraction. Meanwhile, the experimental results also support the effectiveness of our bit significance estimation, which could emphasize informative and spatially coherent RAW responses.

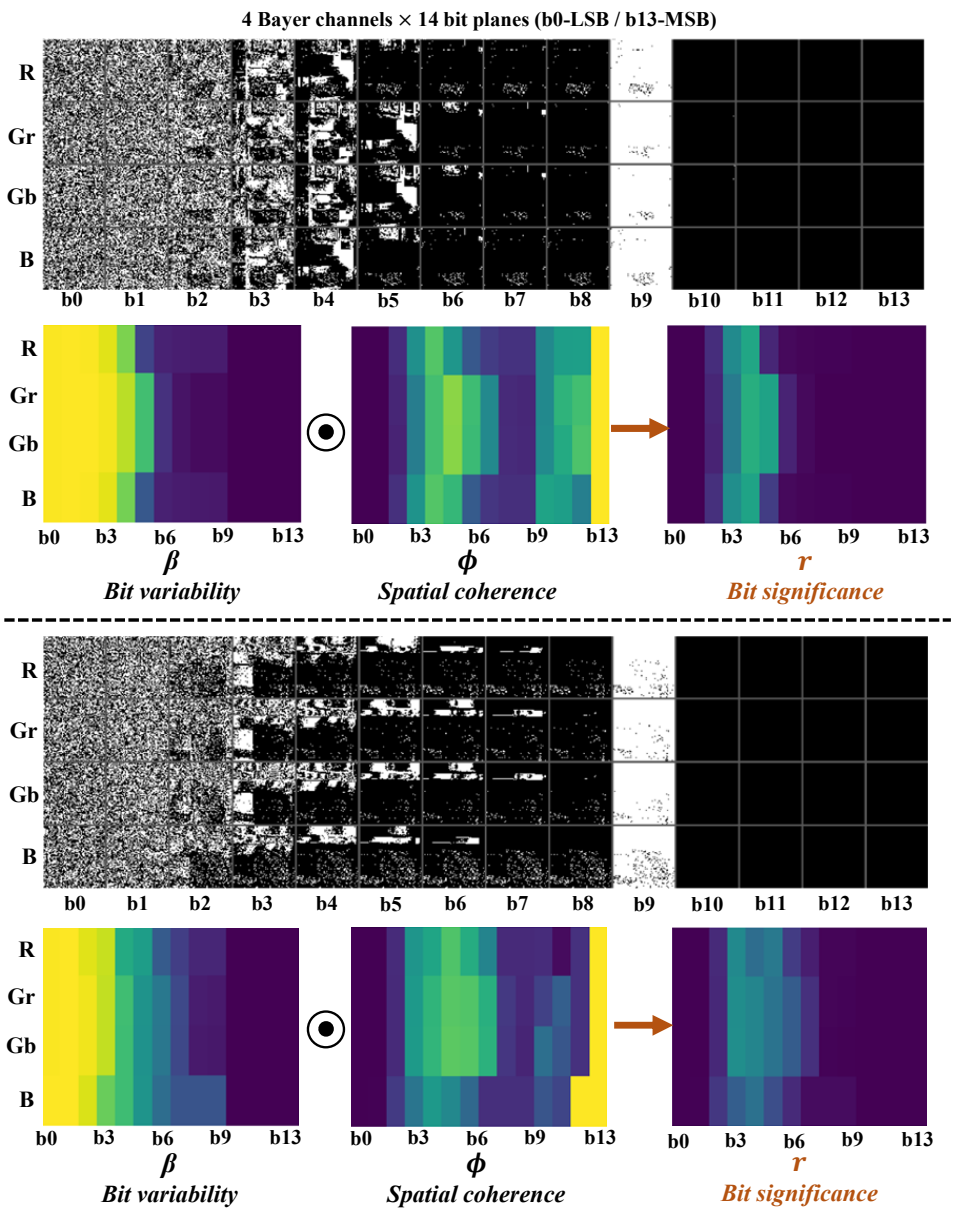



Figure 7: Visualization of bit significance estimation in REC. Two low-light scenes are shown, with binary bit planes and the corresponding bit variability ($\beta$), spatial coherence ($\phi$), and significance ($r = \beta \cdot \phi$) scores for each Bayer channel. The results illustrate the differences across bit positions and Bayer channels, with higher significance scores assigned to bit planes containing informative and spatially coherent structures.

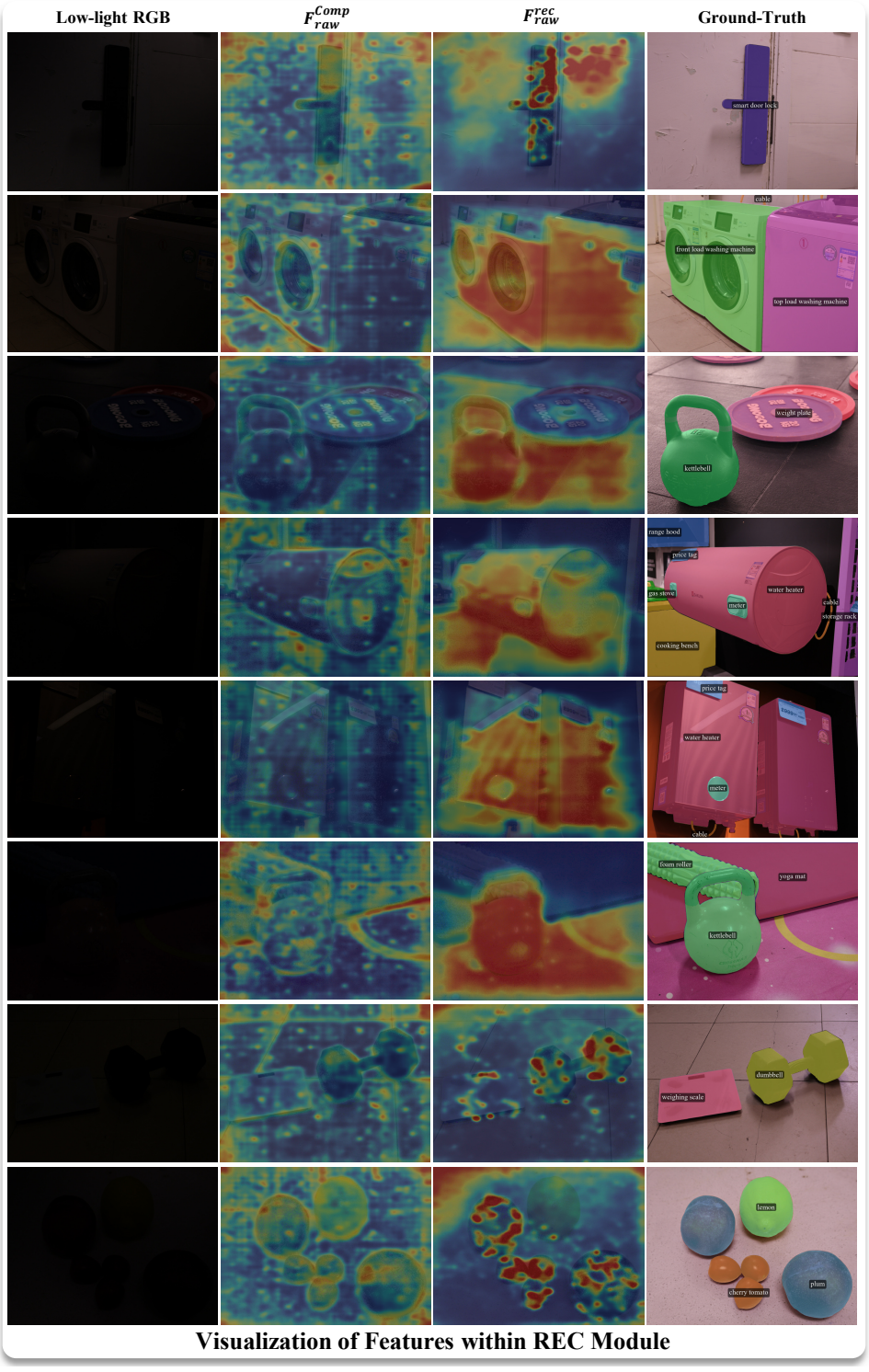



Figure 8: Additional feature visualizations of the REC module. Compared with the directly compressed RAW feature $F_{raw}^{comp}$, the REC-encoded feature $F_{raw}^{real}$ exhibits more concentrated responses on semantically relevant object regions and preserve clearer structural details.




Figure 9: Additional feature visualizations of the RRE module. The local compensation block enhances fine-grained structures, while the global compensation block captures broader spatial dependencies. The final enhanced RGB representation $F^{e}_{rgb}$ exhibits more concentrated responses on semantically relevant object regions.

### A.4 THE EFFECTIVENESS OF REC MODULE

To provide additional evidence for the effectiveness of the REC module, we further visualize more examples of the encoded RAW features in Fig. 8. These results supplement the analysis in Sec. 5.3 of main paper, where only one example is shown due to space limitations. As can be observed, the directly compressed RAW feature $F_{raw}^{comp}$ often exhibits diffused responses and is more easily affected by irrelevant background patterns. In contrast, the REC-encoded feature $F_{raw}^{real}$ produces more concentrated activations on semantically relevant object regions and preserves clearer structural details, such as object boundaries and shapes. This consistent trend can be observed across diverse scenes and object categories. These additional visualizations further support the quantitative results in Table 2, demonstrating that REC can more effectively exploit the informative content in RAW data than direct compression, thereby providing higher-quality RAW features for subsequent RGB representation learning.

### A.5 THE EFFECTIVENESS OF RRE MODULE

We further provide additional visualizations of the intermediate features within the RRE module in Fig. 9, complementing the visualized example shown in the main paper. As illustrated, the original RGB feature $F_{rgb}$ extracted from low-light images often exhibits weak and ambiguous responses, making it difficult to highlight semantically relevant object regions. After the local compensation block, $\mathcal{H}_{local}(rgb)$ enhances fine-grained structures and object boundaries, recovering local visual cues that are weakened under low-light conditions. The subsequent global compensation block, $\mathcal{H}_{global}(local)$, further strengthens object-level responses by modeling broader long-range spatial dependencies. After this progressive enhancement, the final representation $F_{rgb}^{e}$ shows substantially more concentrated responses on the target objects while suppressing irrelevant background activations. Moreover, the resulting feature patterns are qualitatively close to those of the corresponding RAW representation, which is consistent with the quantitative observations in Table 3. These results further demonstrate that the RRE module can effectively lift the degraded RGB features into a more informative RAW-aware space for low-light open-vocabulary segmentation.

## B DETAILS OF DIRECT COMPRESSION

In our experiments, to verify the advantages of the REC module in encoding RAW data, we compare it against a simple direct-compression strategy. As the comparison baseline, we follow common practice in RAW preprocessing (Datta et al., 2023; Li et al., 2025b) and convert the high-bit-depth RAW data into an 8-bit three-channel representation before feature extraction. The detailed pipeline is illustrated in Fig. 10.

Specifically, given a packed RAW image $I_{raw} \in \mathbb{R}^{4\times H/2\times W/2}$ with an RGGB Bayer pattern, we first normalize each Bayer channel into $[0, 1]$:

$$N_{raw}^{c} = \frac{I_{raw}^{c} - x_{\min}}{x_{\max} - x_{\min}}, \qquad c \in \{R, G_r, G_b, B\}, \tag{7}$$

where $I_{raw}^{c}$ denotes the $c$-th Bayer channel, and $x_{\min}$ and $x_{\max}$ denote the minimum and maximum values of the RAW image, respectively. We then merge the $G_r$ and $G_b$ channels by averaging, yielding a three-channel representation:

$$\bar{N}_{raw} = \left[ N_{raw}^{R}, \frac{N_{raw}^{G_r} + N_{raw}^{G_b}}{2}, N_{raw}^{B} \right], \tag{8}$$

where $\bar{N}_{raw} \in \mathbb{R}^{3\times H/2\times W/2}$. Afterwards, bilinear interpolation is applied to restore the spatial resolution to $H \times W$, followed by 8-bit quantization:

$$I_{raw}^{comp} = \mathrm{round}\left(255 \cdot \mathrm{Resize}_{2\times}\left(\bar{N}_{raw}\right)\right), \tag{9}$$

where $I_{raw}^{comp} \in \mathbb{R}^{3\times H\times W}$ denotes the compressed RAW representation in a standard RGB-like format. It is then fed into the RGB visual encoder to obtain $F_{raw}^{comp}$, which serves as the feature representation of the direct-compression baseline in our experiments.

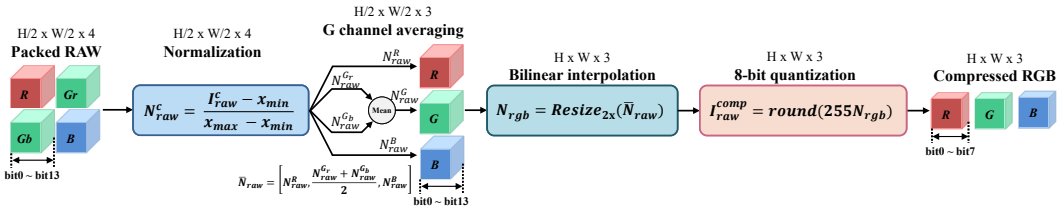



Figure 10: The detailed pipeline of direct-compression.

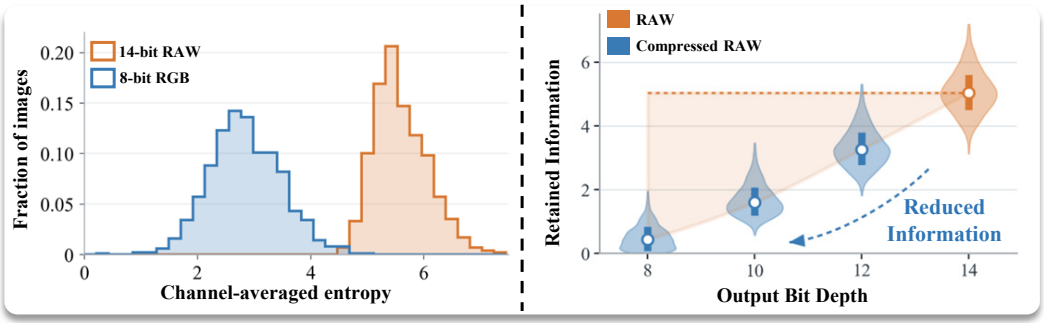



Figure 11: High-bit-depth RAW data preserve richer information than 8-bit RGB images. However, directly compressing RAW data to a lower bit depth leads to substantial information loss.

## C Analysis of Information Retention under Bit-depth Compression

In Fig. 1, we compare the information retained in RAW and RGB images. Meanwhile, we illustrate the information loss caused by directly compressing high-bit-depth RAW data. In this section, we provide the details of the information calculation. For the convenience of the following analysis, we reproduce Fig. 1 here as Fig. 11.

Specifically, for the $c$-th channel of an image $I$ with bit depth $B$, the probability of each intensity value $v$ is computed as:

$$p_c(v) = \frac{\#\{(i,j) \mid I^c(i,j) = v\}}{N_c}, \tag{10}$$

where $N_c$ denotes the number of pixels in the $c$-th channel. The entropy of this channel is then calculated as:

$$H(I^c) = -\sum_{v=0}^{2^B-1} p_c(v) \log_2 p_c(v), \tag{11}$$

where terms with $p_c(v) = 0$ are omitted.

Since packed RAW and RGB images contain different numbers of channels, *i.e.*, four Bayer channels for RAW and three channels for RGB, directly summing the channel-wise entropy would introduce a bias caused by the channel number. We therefore use the channel-averaged entropy for comparison:

$$\bar{H}(I) = \frac{1}{C} \sum_{c=1}^{C} H(I^c), \tag{12}$$

where $C = 4$ for 14-bit RAW and $C = 3$ for 8-bit RGB. As shown in the left part of Fig. 11, the channel-averaged entropy of 14-bit RAW data is consistently higher than that of the corresponding 8-bit RGB images, indicating that high-bit-depth RAW measurements preserve the richer information.

We further investigate the information loss introduced by the compression. Following the direct-compression pipeline described in Sec. B, we compress the RAW data to different bit depths. Notably, we keep all preceding operations in Sec. B unchanged and vary only the final quantization bit

depth. The compressed representation at bit depth $B$ is generated by:

$$I_{raw}^{comp,B} = \text{round}\left[(2^B - 1)\widetilde{N}_{raw}\right], \qquad B \in \{8, 10, 12\}. \tag{13}$$

For each compressed representation, we calculate its channel-averaged entropy using the same formulation above to measure the retained information:

$$R_B = \bar{H}\left(I_{raw}^{comp,B}\right), \qquad B \in \{8, 10, 12\}. \tag{14}$$

For the original 14-bit RAW data, the retained information is directly computed from its four Bayer channels:

$$R_{14} = \bar{H}\left(I_{raw}\right). \tag{15}$$

As shown in the right part of Fig. 11, the retained information progressively decreases as the output bit depth is reduced from 14 to 12, 10, and 8 bits. This result shows that directly compressing high-bit-depth RAW data into a lower-bit-depth representation inevitably causes the information loss. This highlights the importance of our REC module, which explicitly decomposes the original high-bit-depth RAW data into the binary sequence, rather than direct compressing them.

# D The More Details about LOVS Dataset

This section provides additional statistics of the LOVS dataset, including the complete category vocabulary. Meanwhile, we also provide more samples containing paired normal-/low-light images, RAW images, and annotation masks.

## D.1 Category Details

As summarized in Table 4, the category taxonomy of LOVS covers 220 semantic classes spanning 10 different domains, including Living, Food, Cleaning, Office, Retail, Sports, Mobility, Construction, Utilities, and Safety. The complete list of category names for each superclass is also provided in the table. In addition, we also report the number of categories within each superclass and the number of related images. For example, the Retail and Utilities domains appear in 300 and 292 images, respectively, while the Sports domain contains 25 categories appearing in 83 images. Notably, the superclasses are introduced solely for presentation purposes and are not used in training or evaluation. All experiments are conducted at the level of the 220 individual semantic categories. During evaluation, only the 220 annotated LOVS categories are considered. Pixels belonging to all other categories are assigned the ignore label 255 in the ground-truth masks and excluded from the computation of mIoU and hIoU.

Meanwhile, since a single image may contain objects belonging to different classes, the image counts reported for different superclasses are not mutually exclusive, and their sum is larger than the total number of images in LOVS (1,000).

## D.2 Visualization of LOVS Samples

Figure 12 presents additional samples from LOVS, including paired normal-light RGB images, low-light RGB images, corresponding RAW data, and ground-truth annotation masks. The examples cover diverse indoor and outdoor scenes, such as sports equipment, public facilities, construction environments, cleaning supplies, food, and household objects.

Under low-light conditions, object appearance is severely degraded in the RGB images, with weakened colors, textures, and boundaries. The aligned normal-light images provide clearer visual content for reliable annotation, while the paired RAW data retain complementary low-level information from the same scenes. Owing to the aligned acquisition pipeline, the annotations obtained from the normal-light images can be directly transferred to their low-light counterparts. It is worth noting that RAW data are only used during training in our method. During inference, OWL-Net requires only the low-light RGB image to perform open-vocabulary segmentation.

Table 4: The detailed category list of LOVS dataset. The 220 semantic categories are grouped into 10 superclasses, with category names listed alphabetically within each superclass. Image counts indicate the number of images containing at least one category from each superclass.

| Superclass and counts | Category names |
|---|---|
| **Living**<br>36 categories<br>242 images | air conditioner outdoor unit, bath mat, bathroom heater, clothes hanger, clothes peg, countertop, curtain, electric shaver, electric toothbrush, faucet, floor-standing air conditioner, front load washing machine, hair dryer, hand dryer, hand sanitizer, laundry dryer, mirror, remote control, shower, shower loofah, sink, smart door lock, soap dish, statue, table lamp, toothbrush, toothpaste, top load washing machine, towel, umbrella rack, vanity, wall-mounted air conditioner, water dispenser, water heater, water purifier, weighing scale |
| **Food**<br>23 categories<br>131 images | cherry tomato, clay pot, cooking bench, dishwasher, disinfection cabinet, gas stove, induction cooker, juicer, lemon, mango, oven, pear, pineapple, plum, pressure cooker, range hood, refrigerator, rice cooker, steamer pot, strainer bowl, strawberry, thermo jug, watermelon |
| **Cleaning**<br>16 categories<br>127 images | broom, cleaning brush, cleaning glove, cleanser, dustpan, floor scrubber, janitorial cart, laundry detergent, laundry pod, mop, mop bucket, non-recyclable sign, recyclable sign, scouring pad, steel wool, trash can |
| **Office**<br>18 categories<br>125 images | adhesive tape, blackboard, headphone, interactive display, keyboard, laptop, locker, monitor, mouse, mouse pad, podium, printer, projector, projector screen, router, speaker, stapler, whiteboard |
| **Retail**<br>13 categories<br>320 images | display head, parcel, poster, power bank rental station, price tag, self-service terminal, shopping basket, shopping cart, step stool, storage box, storage rack, support stand, vending machine |
| **Sports**<br>25 categories<br>83 images | barbell, basketball, basketball hoop, basketball net, bench press rack, dumbbell, dumbbell rack, foam roller, kettlebell, punching bag, scoreboard, smith machine, soccer, spin bike, sports court, sports court fence, squat rack, stair climber, table tennis net, table tennis table, treadmill, volleyball, weight bench, weight plate, yoga mat |
| **Mobility**<br>30 categories<br>213 images | access control terminal, bike dock station, bollard, charging station, crash barrel, directional board, e-bike windproof quilt, electric two-wheeler, elevator, elevator panel, escalator, flatbed trolley, forklift, guardrail, license plate recognition device, motorhome, parking barrier gate, parking wheel stopper, patrol car, pedestrian turnstile, portable barricade, retractable gate, road sweeper truck, speed bump, sprinkler truck, tactile paving, traffic bollard, traffic cone, traffic sign, tricycle |
| **Construction**<br>17 categories<br>115 images | construction enclosure, electric drill, excavator, fence, hammer, ladder, nail, nut, pliers, screw, screwdriver, shovel, site hoarding, tape measure, tower crane, wheelbarrow, wrench |
| **Utilities**<br>18 categories<br>292 images | battery, cable, cable trunking, distribution box, distribution cabinet, electrical closet, light strip, manhole cover, meter, power bank, power plug, power socket, power strip, power switch, roadside lamp, roller shutter, transmission tower, vent |
| **Safety**<br>24 categories<br>236 images | alarm valve, automated external defibrillator, emergency exit, emergency hammer, emergency light, emergency shelter sign, fire alarm, fire cabinet, fire extinguisher, fire hose, fire hose reel, fire hydrant, fire nozzle, fire sprinkler, fire truck, first aid kit, helmet, security safe, smoke detector, surveillance camera, warning board, warning corner guard, warning light, warning sign |
| **Total** | 220 categories; 1,000 paired normal-/low-light images. |

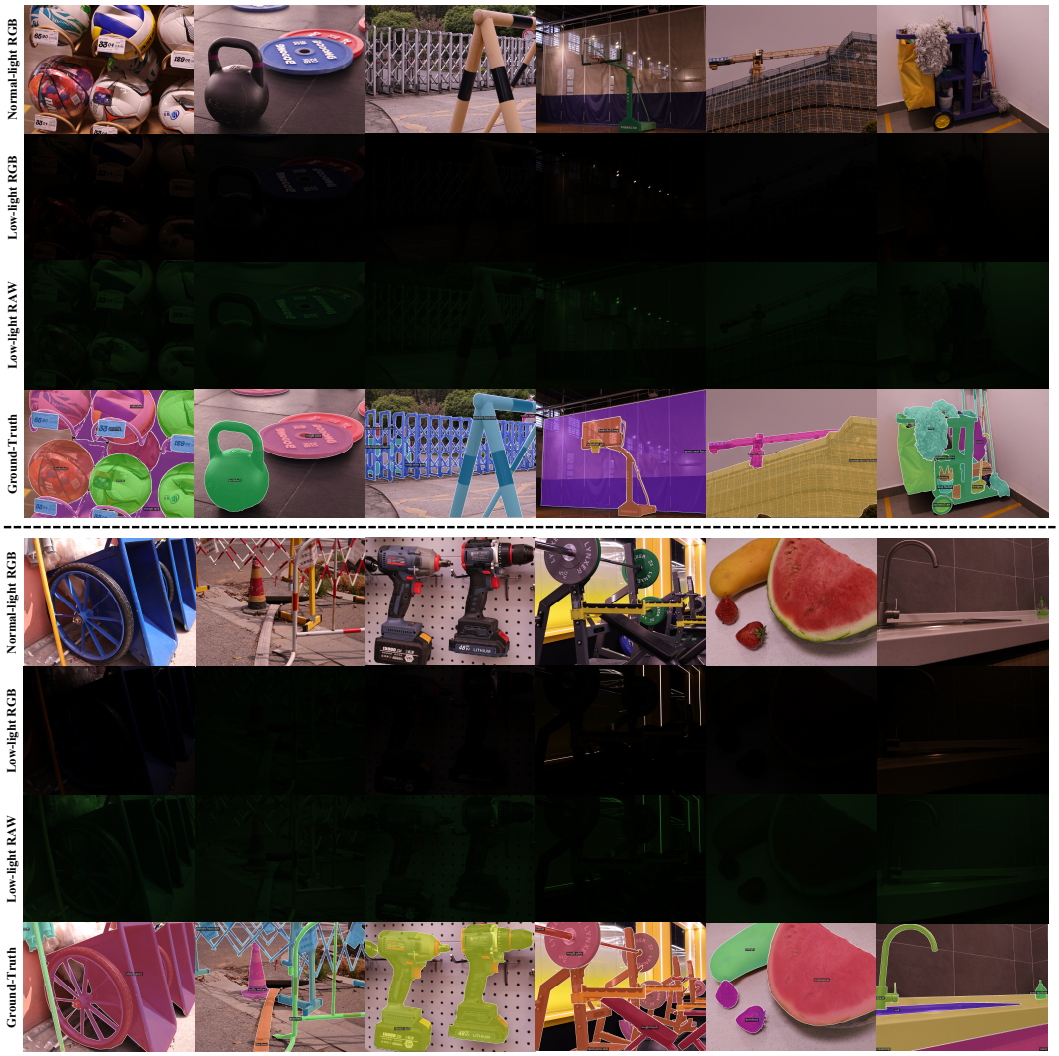



Figure 12: Samples from LOVS. Each group contains four rows, including normal-light RGB images, low-light RGB images, visualized RAW data, and ground-truth segmentation masks overlaid on the normal-light images. Each column corresponds to the same scene across the four rows.

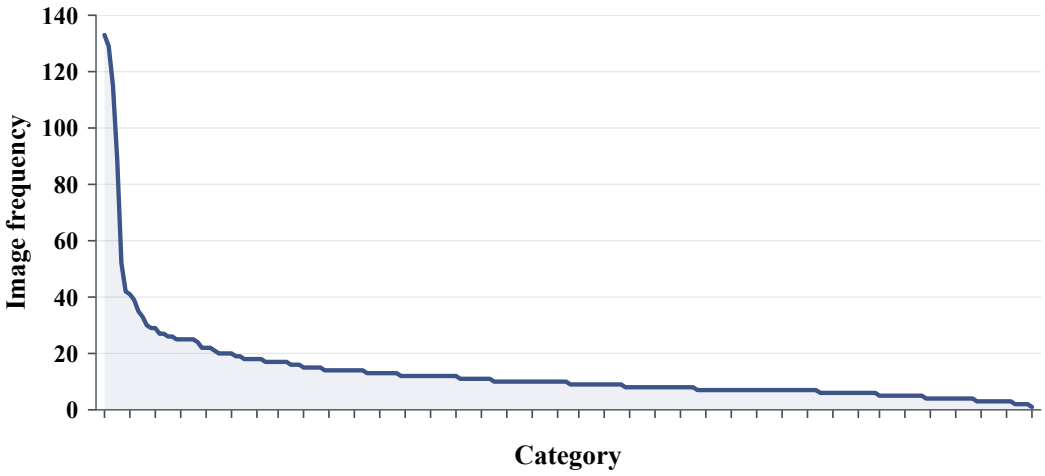



Figure 13: Long-tailed category distribution in LOVS. Categories are sorted by the number of images containing each category, revealing a pronounced long-tailed distribution.

### D.3 Long-tailed Category Distribution

We further analyze the category frequency distribution in LOVS by counting the number of images containing each semantic category. As shown in Fig. 13, LOVS exhibits the long-tailed distribution. Specifically, half of the categories appear in fewer than ten images, while the category frequencies range from only one to 133 images. Such an imbalanced distribution introduces additional challenges for recognizing infrequent categories and further increases the difficulty of open-vocabulary segmentation on LOVS.

## E Implementation Details of Experiments

### E.1 Dark-aware Spatial Fusion Module

To adaptively integrate the enhanced RGB representation $F^e_{rgb}$ with the original RGB feature $F_{rgb}$, we introduce a dark-aware spatial fusion module. Specifically, we first compute an illumination prior $P_i$ from the low-light RGB image as (Jiang et al., 2021):

$$\begin{aligned} Y(x,y) &= 0.299R(x,y) + 0.587G(x,y) + 0.114B(x,y), \\ P_i(x,y) &= \sigma\big(s[\tau - Y(x,y)]\big), \end{aligned} \tag{16}$$

where $Y$ denotes the luminance, $\sigma$ is the sigmoid function, and $s$ and $\tau$ control the sensitivity and threshold of the prior, respectively. A larger value of $P_i$ indicates a darker region where stronger feature enhancement may be needed.

As illustrated in Fig. 2 of the main paper, we first apply cross-attention using $F_{rgb}$ as the query and $F^e_{rgb}$ as the key and value to obtain the feature $C_{fuse}$. $C_{fuse}$ is then concatenated with $F_{rgb}$, which is delivered into two parallel gating branches, producing channel-specific weights at each spatial location and a spatial weight shared across channels. Their outputs are combined with the illumination prior $P_i$ to adaptively modulate $C_{fuse}$, allowing the enhanced representation to complement the original RGB feature according to the local illumination conditions. Finally, the modulated features are added to $F_{\text{rgb}}$ through a residual connection and refined by a feed-forward network with layer normalization, yielding $F_{\text{fuse}}$ for subsequent mask prediction.

### E.2 Details of OWL-Net Across Different Frameworks

In our experiments, we integrate OWL-Net into three representative open-vocabulary segmentation frameworks: CAT-Seg, OVRCOAT, and OpenWorldSAM. Owing to its plug-and-play characteristic,

OWL-Net can be readily incorporated into these frameworks with minimal modifications to their original architectures.

Specifically, for each framework, we leverage the output of visual encoder as the $F_{rgb}$ for subsequent process. Then, the $F_{rgb}$ is enhanced by RRE module to produce the detail recovered representation $F^e_{rgb}$, which is fused with the original RGB embedding $F_{rgb}$ through the dark-aware spatial fusion module. Afterwards, the fused representation $F_{fuse}$ is used as the visual input to the mask decoder for mask prediction. Notably, all the remaining components of the framework are kept unchanged. During the training, we additionally deploy the REC module to extract the RAW features to guide the visual feature learning. During the inference, the REC module is removed, and we only utilize the RGB image as visual input to perform open-vocabulary segmentation.

### E.3 Details of Training Settings

For a fair comparison, we re-train all compared models on the synthetic low-light COCO-Stuff dataset to adapt them to low-light scenarios. Each model is trained using its original hyper-parameter settings on a single NVIDIA H100 GPU. During training, the REC, RRE, and dark-aware spatial fusion modules are trainable, while the visual encoder, text encoder and mask decoder within the original framework remain frozen.

In our experiments, we integrate OWL-Net into three open-vocabulary segmentation frameworks, *i.e.,* CAT-Seg, OVRCOAT, and OpenWorldSAM. For each framework, the RGB-only baseline and its OWL-Net counterpart share identical training settings, including the optimizer, learning-rate schedule, weight decay, batch size, and number of training iterations. All three OWL-Net variants are trained for 80,000 iterations using AdamW, with the learning rate reduced by a factor of 10 at 64,000 and 72,000 iterations.

**OpenWorldSAM.** We train OpenWorldSAM and its OWL-Net counterpart with an initial learning rate of $1 \times 10^{-4}$, a weight decay of $0.05$, and a total batch size of 8.

**CAT-Seg.** We train CAT-Seg and its OWL-Net counterpart with an initial learning rate of $1 \times 10^{-4}$, a weight decay of $1 \times 10^{-4}$, and a total batch size of 4.

**OVRCOAT.** We train OVRCOAT and its OWL-Net counterpart with an initial learning rate of $5 \times 10^{-5}$, a weight decay of $0.05$, and a total batch size of 4.

Table 5: Comparison of segmentation accuracy and inference efficiency on LOVS. Two-stage methods first enhance the input degraded images and then perform open-vocabulary segmentation. In contrast, OWL-Net directly processes low-light RGB images. As shown, our method achieves competitive performance with substantially higher inference efficiency.

| Method | mIoU↑ | hIoU↑ | Mean time (ms)↓ | FPS↑ | Mean GFLOPs↓ |
|---|---|---|---|---|---|
| OpenWorldSAM | 48.1 | 46.0 | 126.04 | 7.93 | 3,132.40 |
| HVI-CIDNet + OpenWorldSAM | 51.0 | 48.1 | 386.24 | 2.59 | 3,464.00 |
| Reti-Diff + OpenWorldSAM | 52.9 | 50.3 | 1,552.41 | 0.64 | 12,454.82 |
| OWL-Net(OpenWorldSAM) | 49.2 | 51.3 | 132.55 | 7.54 | 3,254.83 |

## F Comparison with Two-Stage Low-Light Open-Vocabulary Segmentation

To further demonstrate the effectiveness and efficiency of OWL-Net, we compare it with two-stage approaches that first enhance low-light images and then perform open-vocabulary segmentation. We evaluate each method on LOVS from two aspects, *i.e.,* segmentation accuracy and inference efficiency. All experiments are conducted under the same hardware settings.

As shown in Table 5, the two-stage methods require an additional enhancement model to recover visual details before segmentation, therefore introducing considerable computational overhead. In contrast, OWL-Net directly segments low-light RGB images without explicit image enhancement, achieving 49.2% mIoU and 51.3% hIoU with an inference time of only 132.55 ms per image.

Specifically, HVI-CIDNet (Yan et al., 2025) + OpenWorldSAM and Reti-Diff (He et al., 2025) + OpenWorldSAM require 386.24 ms and 1,552.41 ms to process an image, respectively. Therefore, OWL-Net is approximately $2.9\times$ and $11.7\times$ faster than these two-stage approaches, respectively, while maintaining competitive segmentation accuracy.

Moreover, we compare OWL-Net with its corresponding baseline, OpenWorldSAM. As reported in Table 5, OWL-Net introduces only marginal computational overhead and maintains a comparable inference speed. In contrast, the segmentation performance is clearly improved, with mIoU increasing from 48.1% to 49.2% and hIoU from 46.0% to 51.3%. These results demonstrate that OWL-Net substantially improves low-light open-vocabulary segmentation accuracy with only a minor increase in computational cost, further demonstrating the superiority of our method.

Table 6: Mean RGB-to-RAW feature alignment during cosine distillation. Lower distances and higher cosine similarity indicate closer alignment.

| Iterations | Mean L1 ↓ | Mean L2 ↓ | Mean MSE ↓ | Mean Cosine ↑ |
|---|---|---|---|---|
| 0 | 1.130 | 22.725 | 2.023 | -0.012 |
| 50 | 1.119 | 22.477 | 1.979 | 0.010 |
| 100 | 1.041 | 20.799 | 1.695 | 0.153 |
| 200 | 0.797 | 15.939 | 0.996 | 0.504 |
| 500 | 0.429 | 8.629 | 0.293 | 0.856 |
| 1,000 | 0.366 | 7.367 | 0.218 | 0.895 |
| 2,000 | 0.436 | 8.683 | 0.323 | 0.848 |
| 5,000 | 0.318 | 6.363 | 0.181 | 0.917 |

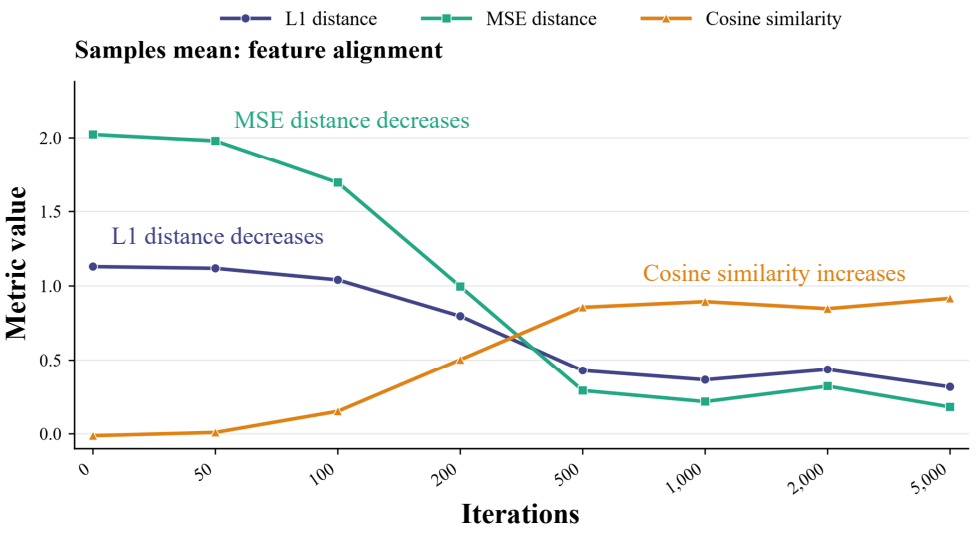



Figure 14: Evolution of RGB-to-RAW feature alignment during training. The decreasing L1 and MSE distances and increasing cosine similarity indicate that the enhanced RGB features progressively approach the real RAW representations.

## G The Evolution of Enhanced RGB Representation

To verify that the RRE module indeed lifts the RGB features toward a RAW-consistent space, rather than merely learning performance-friendly latent features, we track the alignment between the enhanced RGB feature $F^{e}_{rgb}$ and the real RAW feature $F^{real}_{raw}$ at different training iterations. Specifically, we measure their L1 distance, L2 distance, mean squared error (MSE), and cosine similarity throughout training. Lower distances and higher cosine similarity indicate a closer correspondence between the two representations.

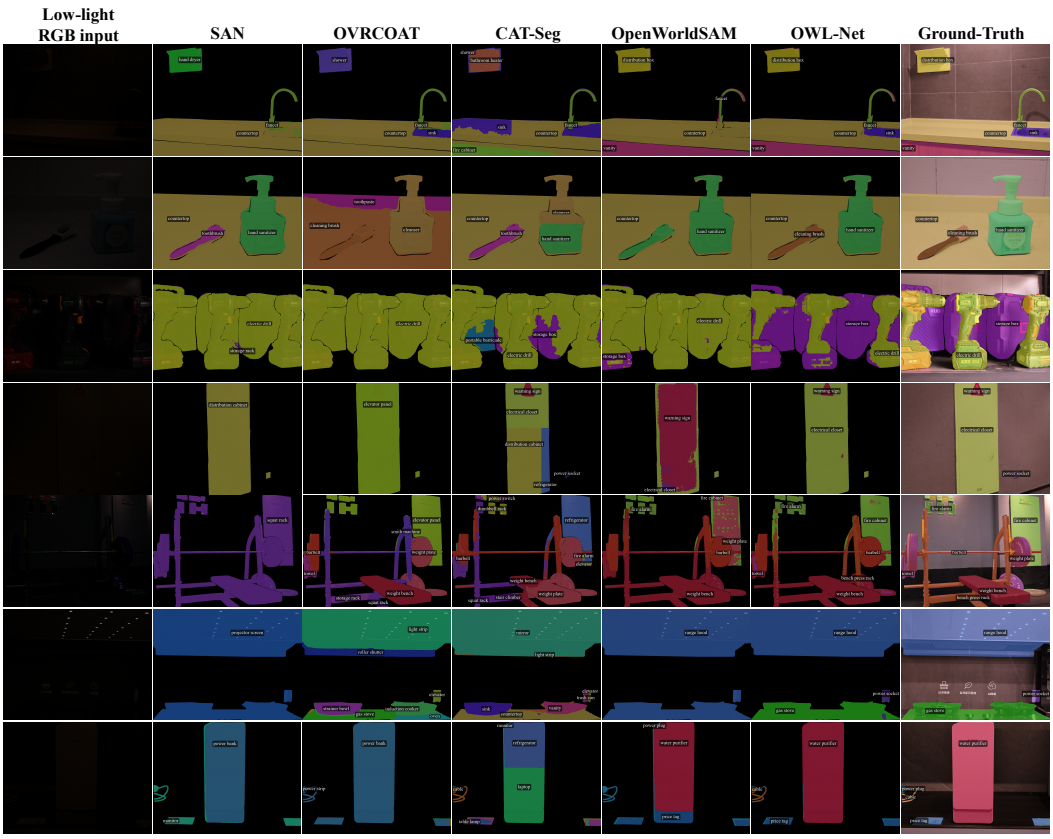



Figure 15: We visualize segmentation maps from different open-vocabulary segmentation methods under low-light conditions. As shown, OWL-Net produces more complete object regions and more accurate segmentation boundaries, achieving better segmentation results under low-light conditions.

As reported in Table 6 and Fig. 14, the enhanced RGB representation progressively approaches the real RAW representation as training proceeds. The mean L1 distance decreases from 1.130 at initialization to 0.318 at 5,000 iterations, while the MSE decreases from 2.023 to 0.181. Meanwhile, the cosine similarity increases substantially from $-0.012$ to 0.917. These results demonstrate that RRE progressively learns to produce informative representations from degraded RGB features, which closely approximate those extracted from real RAW data.

## H Additional Qualitative Comparisons

In Fig. 15, we provide additional qualitative comparisons of different open-vocabulary segmentation methods on LOVS. Compared with existing methods, OWL-Net produces more complete object regions and more accurate segmentation boundaries under low-light conditions, further demonstrating the effectiveness of OWL-Net in challenging low-light scenarios.

## I Evaluation Metrics

We evaluate segmentation performance on LOVS using mean Intersection over Union (mIoU) and harmonic mean Intersection over Union (hIoU). For each semantic category $c$, the Intersection over Union (IoU) is defined as:

$$\mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \tag{17}$$

where $\mathrm{TP}_c$, $\mathrm{FP}_c$, and $\mathrm{FN}_c$ denote the total numbers of true-positive, false-positive, and false-negative pixels for category $c$, respectively. Then, the mean IoU is computed by averaging the

category-wise IoU over all $C$ evaluation classes:

$$\mathrm{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\mathrm{IoU}_c. \tag{18}$$

To further evaluate the performance on seen and unseen categories, we report hIoU in our experiments. Since COCO-Stuff covers a broad range of thing and stuff categories, semantic overlap between its training set and downstream open-vocabulary segmentation benchmarks is common. For example, previous work (Xu et al., 2023) has shown that standard evaluation datasets in this domain, including PASCAL VOC, PASCAL Context, and ADE20K, exhibit varying degrees of label-set overlap with COCO-Stuff. In LOVS, 14 out of the 220 semantic categories are also present in the COCO-Stuff training label space, including *sink, towel, hair dryer, fire hydrant, laptop, toothbrush, fence, mouse, keyboard, oven, curtain, refrigerator, remote control, mirror*. We regard these 14 categories as seen categories and the remaining 206 categories as unseen categories. Let $\mathrm{mIoU}_s$ and $\mathrm{mIoU}_u$ denote the mean IoU over the seen and unseen categories, respectively. hIoU is computed as their harmonic mean:

$$\mathrm{hIoU} = \frac{2\,\mathrm{mIoU}_s\,\mathrm{mIoU}_u}{\mathrm{mIoU}_s + \mathrm{mIoU}_u}. \tag{19}$$

A high hIoU requires strong performance on both seen and unseen categories and therefore provides a balanced measure of open-vocabulary segmentation performance. In particular, mIoU alone may obscure an imbalanced performance across the two subsets: a model performing well on seen categories but poorly on unseen ones can still obtain a relatively high overall mIoU. By taking the harmonic mean of $\mathrm{mIoU}_s$ and $\mathrm{mIoU}_u$, hIoU penalizes such imbalance and better reflects the model's ability to generalize to unseen categories.

## J Effect of Different Feature Alignment Losses

To investigate the effect of different loss functions on the RGB Representation Enhancement (RRE) module, we compare four feature alignment losses: mean squared error (MSE), L2 distance, Kullback–Leibler (KL) divergence, and cosine similarity loss. All four losses are used to align the enhanced RGB features $F^{e}_{rgb}$ with the corresponding RAW features $F^{real}_{raw}$, while keeping the remaining experimental settings unchanged.

As shown in Table 7, cosine similarity loss achieves the highest mIoU and hIoU on LOVS, reaching 49.2% and 51.3%, respectively. These results indicate that cosine similarity is an effective choice for aligning the enhanced RGB features with the RAW features in our framework. We therefore adopt cosine similarity loss in the RRE module. Meanwhile, replacing cosine similarity with other alignment losses only results in slight performance degradation, indicating that our framework is relatively robust to the specific choice of feature alignment loss.

Table 7: Effect of different feature alignment losses on LOVS. Each loss aligns the enhanced RGB features $F^{e}_{rgb}$ with the corresponding RAW features $F^{real}_{raw}$.

| Alignment Loss | mIoU↑ | hIoU↑ |
|---|---|---|
| MSE | 48.6 | 49.7 |
| L2 distance | 48.9 | 48.2 |
| KL divergence | 48.1 | 49.1 |
| Cosine similarity | **49.2** | **51.3** |

## K Synthesis of Low-Light RAW and RGB Images

Due to the lack of low-light images with open-vocabulary segmentation annotations, we synthesize paired low-light RGB and RAW data for training. Specifically, we use normal-light images from COCO-Stuff as the source data, based on which the corresponding low-light RAW data with RGB images are generated. Meanwhile, we preserve the original segmentation annotations within

COCO-Stuff. Following prior work (Chen et al., 2023), we adopt the Physics-based Noise Formation model (Wei et al., 2020) to synthesize realistic low-light degradation.

Specifically, we first convert each normal-light RGB image into synthetic Bayer RAW data through an approximate inverse image signal processing (ISP) pipeline. We then scale the RAW intensities by 0.05 and add sensor noise using the Canon EOS 5D Mark IV calibration profile, with the noise ratio set to 5. This produces synthetic low-light RAW data. To generate the corresponding low-light RGB image, we apply channel-wise bilateral denoising ($\sigma_s = 1.2$, $\sigma_c = 0.2$), followed by a forward ISP using the color correction matrix and gains from the inverse pipeline.

To assess synthesis fidelity, we apply the above procedure to the normal-light RGB images in LOVS and LIS (Chen et al., 2023). Then, we compare the synthesized outputs with their corresponding real low-light RGB images within each dataset. As shown in Table 8, the synthetic images achieve PSNR values of 34.0 dB on LOVS and 32.2 dB on LIS. These results indicate close agreement between synthetic low-light images and the real ones, supporting the fidelity of our synthesis pipeline.

Table 8: PSNR comparison between synthesized low-light RGB images and their corresponding real ones on LOVS and LIS. The high similarity on both datasets demonstrates the fidelity of the synthesis pipeline utilized in our method.

| Dataset | PSNR (dB)↑ |
|---|---|
| LOVS | 34.0 |
| LIS | 32.2 |