Saliency–Depth Conditioning for Zero-Shot Segmentation of Communication-Tower Components in Cluttered UAV Imagery

Ali Lesani[1,2], **Chul Min Yeum**[1,2], **Su-Min Kang**[3]

[1] Computer Vision for Smart Structures (CViSS) Lab, Waterloo, Canada
[2] University of Waterloo, Waterloo, Canada
[3] School of Architecture, Soongsil University, Seoul, South Korea

Abstract

Fine-grained segmentation of communication-tower components in UAV imagery is essential for automated inspection, yet task-specific models are hard to develop due to limited instance-level annotations. Zero-shot segmentation models offer a promising alternative, but in cluttered scenes, visually similar background structures interfere with component localization, causing missed instances and false positives. We propose a model-agnostic saliency–depth foreground-conditioning strategy combining appearance-based saliency with monocular relative depth to construct a coarse tower prior and suppress irrelevant content. We integrate this module with Grounded-SAM and SAM 3, yielding SD-Grounded-SAM and SD-SAM 3. SD-Grounded-SAM further applies geometric and depth-aware box refinement before mask generation, while SD-SAM 3 relies on SAM 3's internal setup. On TOW-300, a dataset of 340 communication-tower UAV images, our strategy improves both baselines: SD-SAM 3 achieves the strongest instance-segmentation performance, while SD-Grounded-SAM produces fewer false positives. Ablations confirm complementary gains from saliency, depth, and box refinement, improving robustness in cluttered scenes.

1 Introduction

Communication towers are critical components of modern wireless infrastructure, enabling large-scale cellular connectivity. Accurate visual understanding of these structures—particularly the ability to identify and segment individual components such as antennas and radios—is essential for downstream applications including asset inventory, automated maintenance, change detection, structural assessment, anomaly detection, and network optimization [1, 2, 3]. As these tasks increasingly rely on high-resolution visual data collected in the field, unmanned aerial vehicle (UAV)-based inspection has emerged as a practical solution, offering a safe, efficient, and cost-effective alternative to manual climbing or ground-based imaging. UAVs can capture comprehensive visual coverage of towers from multiple viewpoints and elevations, leading to widespread adoption in infrastructure monitoring [1, 2, 4, 5]. However, the resulting images are often visually complex and contain substantial background clutter, making automated component-level segmentation challenging.

This challenge is further amplified by variations introduced during UAV image acquisition and by the visual characteristics of tower-mounted components. UAV flight paths produce considerable changes in viewpoint, scale, elevation, and perspective, causing the same component to appear differently across images. In addition, antennas and radio units are often small relative to the full-resolution image, overlap with the tower structure, or are partially occluded by neighboring equipment. Their localization is further complicated by surrounding vegetation, buildings, cables, power lines, and metallic or rectangular structures that may exhibit similar visual patterns. Related difficulties have

also been reported in UAV-based inspection of power-transmission infrastructure, particularly for detecting small components under cluttered backgrounds and large scale variations [2, 4, 6, 7, 8]. Consequently, successful component segmentation depends not only on accurate mask generation, but also on reliable localization of small target structures in visually crowded scenes.

Addressing these localization and segmentation challenges has traditionally relied on supervised instance segmentation models [9, 10, 11]. These models can achieve strong performance when sufficient task-specific annotations are available. However, constructing large instance-level datasets for communication-tower components is costly and labor-intensive because each small, overlapping, or partially occluded component must be annotated at the pixel level. Moreover, models trained on a fixed dataset may not generalize reliably to changes in tower configuration, antenna design, image-acquisition conditions, and surrounding environments. Data scarcity and distribution shift therefore remain major barriers to the practical deployment of supervised segmentation models in infrastructure inspection [1, 2, 5]. These limitations motivate the investigation of training-free approaches that leverage general-purpose pre-trained models without requiring a dedicated segmentation model for each tower configuration.

Recent foundation models provide new opportunities for prompt-based and zero-shot segmentation. The Segment Anything Model (SAM) can generate high-quality masks from user-provided prompts, such as points or bounding boxes, and has demonstrated strong transfer across diverse image distributions [12]. However, SAM does not independently localize objects of interest and therefore depends on accurate prompts to produce meaningful masks. In large-scale UAV inspection, manually specifying prompts for every antenna or radio unit is impractical. Grounded-SAM addresses this limitation by combining the open-vocabulary detector Grounding DINO with SAM, allowing bounding-box prompts to be generated automatically from natural-language descriptions [13, 14]. More recent concept-prompted models, such as SAM 3, further unify object localization and mask generation within a single zero-shot segmentation framework.

Despite these advances, open-vocabulary localization remains difficult in communication-tower imagery. Object names such as *antenna* are ambiguous because the detector has limited domain-specific knowledge of tower-mounted equipment. In our observations, such prompts may identify the complete tower, structural members, or unrelated objects rather than individual mounted components. More descriptive appearance-based prompts can improve localization of antenna-like regions, but they also increase false detections on background elements with similar geometry. These errors arise during object localization and can propagate to the final masks, even when the underlying segmentation model is capable of accurately delineating a prompted region. Therefore, the central challenge is not only mask generation, but also reducing the visual ambiguity presented to the zero-shot model.

This limitation motivates the use of visual pre-processing before prompt-based detection and segmentation. In particular, reducing irrelevant background content can remove many non-target objects that satisfy coarse geometric text descriptions and thereby reduce ambiguity for downstream zero-shot models. To this end, we introduce a model-agnostic saliency–depth foreground-conditioning strategy for training-free segmentation of communication-tower components in UAV imagery. The proposed module combines appearance-based saliency with monocular relative depth to construct a coarse tower prior. Saliency-based background suppression first isolates visually prominent foreground regions and reduces the search space presented to subsequent models [15]. However, saliency estimation may remove low-contrast, partially occluded, or visually weak portions of the tower. We therefore incorporate monocular depth estimation, exploiting the observation that the communication tower is generally closer to the UAV camera than much of the surrounding background [16]. Depth cues are used to recover near-field regions that may have been omitted by saliency estimation. Morphological refinement and connected-component filtering are then applied to produce a coherent foreground prior, which is used to construct a conditioned image with reduced background interference.

We integrate the proposed conditioning strategy with two distinct zero-shot segmentation frameworks. First, we develop Saliency–Depth Grounded-SAM (SD-Grounded-SAM), a domain-aware extension of Grounded-SAM. In this configuration, Grounding DINO operates on the conditioned image to generate candidate boxes for tower-mounted components. The candidate detections are subsequently refined using tower-relative geometry, pairwise containment, and relative-depth cues. The refined boxes are then provided to SAM as spatial prompts for instance-mask generation on the original image. Second, we apply the same saliency–depth conditioning module to SAM 3, resulting in Saliency–

Depth SAM 3 (SD-SAM 3). In contrast to Grounded-SAM, SAM 3 performs localization and segmentation internally; therefore, the conditioned image is passed directly to its concept-prompted inference pipeline without an explicit box-refinement stage. These two integrations allow us to evaluate whether the proposed foreground-conditioning strategy improves architecturally distinct zero-shot segmentation systems.

We evaluate the proposed method using TOW-300, a newly developed dataset comprising 340 high-resolution UAV images collected under real-world field conditions, with instance-level annotations of tower-mounted components. The saliency–depth conditioning strategy consistently improves both Grounded-SAM and SAM 3 across instance- and image-level segmentation metrics. The SD-SAM 3 configuration achieves the strongest overall instance segmentation performance, while the full SD-Grounded-SAM pipeline provides a more conservative operating point with higher precision and cleaner aggregate foreground masks.

We summarize our primary contributions as follows: (1) we introduce a model-agnostic and training-free saliency–depth foreground-conditioning strategy that combines complementary appearance and geometric cues to reduce background ambiguity in cluttered UAV imagery; (2) we integrate the proposed conditioning strategy with two distinct zero-shot segmentation frameworks, resulting in SD-Grounded-SAM and SD-SAM 3, and demonstrate consistent improvements over their corresponding unconditioned baselines; (3) for the Grounded-SAM-based integration, we develop an additional box-refinement stage that uses tower-relative geometry, pairwise containment, and relative-depth cues to produce more reliable spatial prompts for SAM; and (4) we conduct a comprehensive evaluation on real-world communication-tower imagery, showing that the proposed conditioning strategy improves instance localization, recall, mask overlap, and aggregate foreground quality across both segmentation frameworks.

2 Related Work

UAV-Based Visual Inspection. UAV-based visual inspection is widely used to collect visual data from infrastructure that is difficult, costly, or unsafe to access manually [17, 18, 19]. Recent reviews show that UAVs are used across a broad range of infrastructure applications, including bridges, buildings, transportation assets, and utility systems, because they enable flexible image acquisition and support efficient downstream analysis [18, 20]. In this workflow, deep learning has become an important tool for automating the detection, classification, and segmentation of defects and structural components from UAV imagery [1, 21, 22, 23, 24]. In aerial infrastructure vision, closely related work has focused on transmission towers, power lines, and insulators [25, 26, 27, 26]. For example, TTPLA introduced a benchmark dataset for detection and segmentation of transmission towers and power lines in aerial images [28], while other studies have developed UAV-based methods for insulator detection under complex backgrounds and large scale variation [29, 2]. These studies establish UAV inspection as a practical computer vision setting, but they mainly address power infrastructure or defect detection rather than fine-grained segmentation of communication-tower antennas and radios.

Supervised Segmentation. Supervised segmentation methods remain the dominant approach for dense visual understanding. Architectures such as U-Net and DeepLabv3+ have become standard baselines for semantic segmentation, while models such as Mask R-CNN, YOLACT, and SOLOv2 are widely used for instance segmentation across domain-specific vision tasks [30, 31, 10, 9, 11, 32, 33]. In infrastructure inspection, supervised models have also been adapted to aerial tower and power-line imagery, often with task-specific datasets and closed-set label definitions [28, 34, 29]. While these methods can achieve strong results when sufficient labeled data are available, they are less attractive in settings where fine-grained annotation is expensive and target components vary across hardware types and deployment environments. Our work differs from this line by addressing communication-tower component segmentation in a zero-shot setting without training a dedicated supervised segmenter.

Foundation Models for Zero-Shot Segmentation. Recent foundation models have significantly advanced zero-shot and open-vocabulary segmentation by leveraging large-scale visual and vision–language pretraining. Methods such as CLIPSeg, LSeg, and GroupViT explore language-guided segmentation and open-vocabulary recognition across diverse image domains [35, 36, 37]. SAM further introduced large-scale promptable segmentation and demonstrated strong transfer to previously unseen image distributions [12]. Building on SAM, Grounded-SAM combines open-vocabulary detection with promptable mask generation to enable text-guided instance segmentation [13, 14], while

more recent models such as SAM 3 integrate concept-based localization, segmentation, and tracking within a unified framework [38]. Despite their broad generalization capabilities, these models are designed as general-purpose vision systems rather than being tailored to the structural characteristics of infrastructure-inspection imagery [39, 40]. Existing approaches primarily improve segmentation through stronger prompting, open-vocabulary localization, or unified detection–segmentation architectures, whereas comparatively less attention has been given to modifying the visual input itself to reduce domain-specific scene ambiguity. Our work addresses this gap by introducing a model-agnostic foreground-conditioning stage based on complementary saliency and depth cues. Rather than modifying or fine-tuning the underlying zero-shot models, the proposed strategy restructures the visual evidence presented to them, allowing both Grounded-SAM and SAM 3 to operate on a reduced and more relevant search space.

3 Method

3.1 Problem Setup

In this work, we address the problem of zero-shot instance-level segmentation of communication-tower components from UAV-captured RGB images. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, our goal is to generate a set of instance masks

$$\mathcal{M} = \{\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_K\}, \tag{1}$$

where K denotes the number of predicted component instances, and $\mathbf{M}_k \in \{0,1\}^{H \times W}$ denotes the binary mask of the kth predicted instance, with $k \in \{1, \ldots, K\}$. Each mask corresponds to an individual tower-mounted component, such as an antenna or radio unit.

The objective is to produce accurate instance-level masks without task-specific model training. Instead of training a dedicated segmentation network, we use general-purpose pre-trained models and introduce a shared saliency–depth foreground-conditioning module that reduces background ambiguity before zero-shot segmentation. Given the input image, the conditioning module produces a refined foreground image

$$\mathbf{I}_{\mathrm{ref}} = \mathcal{R}(\mathbf{I}), \tag{2}$$

where $\mathcal{R}(\cdot)$ denotes the proposed saliency–depth conditioning operation. The resulting conditioned image $\mathbf{I}_{\mathrm{ref}}$, together with a textual prompt $\mathbf{p}$ describing the target components, is then provided to one of two zero-shot segmentation frameworks:

$$\mathcal{M}^{(j)} = f_j\left(\mathbf{I}, \mathbf{I}_{\mathrm{ref}}, \mathbf{p}\right), \qquad j \in \{\mathrm{GSAM}, \mathrm{SAM3}\}, \tag{3}$$

where $f_{\mathrm{GSAM}}(\cdot)$ denotes the SD-Grounded-SAM pipeline and $f_{\mathrm{SAM3}}(\cdot)$ denotes the SD-SAM 3 pipeline.

3.2 Saliency–Depth Foreground Conditioning

An overview of the proposed framework is shown in Figure 1. Starting from a raw RGB image captured by a UAV, the conditioning module first estimates visually salient foreground regions and predicts a monocular relative-depth map. The two cues are then combined to construct a coarse tower prior. Morphological refinement and connected-component filtering are applied to remove isolated noise and obtain a coherent foreground region. The resulting mask is used to generate a conditioned image that preserves the dominant tower structure while suppressing much of the surrounding background. The conditioning module is shared by both downstream zero-shot segmentation frameworks. Its purpose is not to perform final component segmentation directly, but to reduce the visual ambiguity presented to the segmentation model.

Saliency-Based Foreground Estimation. To reduce the effect of background clutter, we first estimate a saliency map from the input image. We use the Transparent Background model, which is based on a deep salient-object-detection network that separates visually prominent foreground regions from the background [15]. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, the model produces a saliency map

$$\mathbf{S} \in [0,1]^{H \times W}, \tag{4}$$

where larger values indicate a higher likelihood of belonging to the salient foreground. The saliency map is converted into an initial binary foreground mask:

$$\mathbf{M}_{\mathrm{sal}}(x,y) = \mathbb{1}\left[\mathbf{S}(x,y) \geq \tau_{\mathrm{sal}}\right], \tag{5}$$

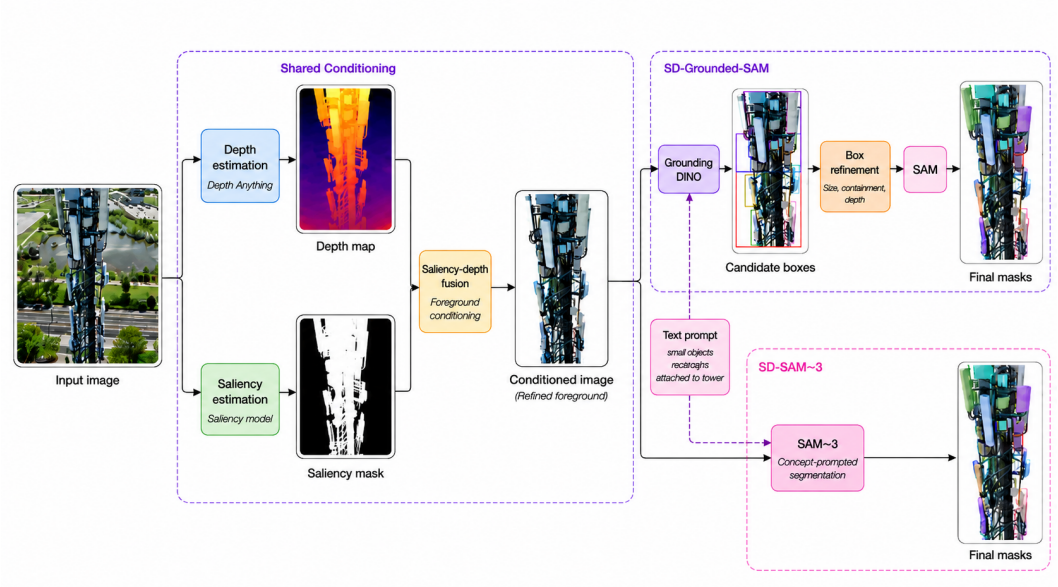

Figure 1: **Overview of the proposed saliency–depth conditioning framework.** Given an input UAV image, saliency estimation and monocular depth prediction are combined to construct a refined foreground mask and a conditioned image with reduced background interference. The conditioned image is then processed by two zero-shot segmentation frameworks. In SD-Grounded-SAM, Grounding DINO generates candidate boxes that are refined using tower-relative geometry, pairwise containment, and depth cues before being passed to SAM. In SD-SAM 3, the conditioned image and text prompt are provided directly to SAM 3, which performs localization and segmentation internally.

where τ_{sal} is the saliency threshold and $\mathbb{1}[\cdot]$ denotes the indicator function. Applying this mask suppresses large portions of the irrelevant background and produces an initial foreground estimate focused primarily on the tower structure. However, saliency-based methods rely mainly on appearance cues and may fail under low contrast, occlusion, or complex textures. As a result, portions of the tower may be incorrectly suppressed as background. We therefore refine the foreground estimate using monocular depth information.

Depth-Based Foreground Refinement. To recover relevant regions that may be removed during saliency estimation, we incorporate monocular depth prediction using Depth Anything [16]. Given the same input image $\mathbf{I}$, the model produces a relative depth map

$$\mathbf{D} \in \mathbb{R}^{H \times W}. \tag{6}$$

Based on the observation that the communication tower and its mounted components are typically closer to the UAV camera than much of the surrounding background, the depth map provides a complementary geometric cue for foreground identification. We construct a binary near-field recovery mask as

$$\mathbf{M}_{\text{dep}}(x, y) = \mathbb{1}\left[\mathbf{D}(x, y) \geq \tau_{\text{rec}}\right], \tag{7}$$

where $\tau_{\text{rec}} \in [0, 255]$ is the depth threshold used to identify near-field pixels. In the employed depth representation, larger values correspond to regions estimated to be closer to the camera.

Saliency–Depth Mask Fusion and Foreground Construction. The saliency and depth masks are first combined through a pixel-wise union:

$$\mathbf{M}_0 = \mathbf{M}_{\text{sal}} \vee \mathbf{M}_{\text{dep}}, \tag{8}$$

where $\vee$ denotes the pixel-wise logical OR operation. This operation preserves regions identified as salient while reintroducing near-field regions that may have been incorrectly removed by the saliency model.

The combined mask may still contain small holes, disconnected regions, and isolated noise. We therefore refine it using morphological closing followed by morphological opening:

$$\widetilde{\mathbf{M}} = \mathrm{Open}_{\mathcal{K}_5}\left(\mathrm{Close}_{\mathcal{K}_9}\left(\mathbf{M}_0\right)\right), \tag{9}$$

where $\mathcal{K}_9$ and $\mathcal{K}_5$ denote 9×9 and 5×5 square structuring elements, respectively. Morphological closing fills small gaps and connects nearby foreground regions, whereas morphological opening removes isolated noise and small spurious structures.

We then retain the largest connected component in the morphologically refined mask:

$$\mathbf{M}_{\mathrm{ref}} = \mathrm{LCC}_8\left(\widetilde{\mathbf{M}}\right), \tag{10}$$

where $\mathrm{LCC}_8(\cdot)$ denotes selection of the largest connected component using 8-neighborhood connectivity. This step removes remaining disconnected regions under the assumption that the tower forms the dominant foreground structure.

The conditioned foreground image is obtained by applying the resulting mask to the original RGB image:

$$\mathbf{I}_{\mathrm{ref}} = \mathbf{I} \odot \mathbf{M}_{\mathrm{ref}}, \tag{11}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{M}_{\mathrm{ref}}$ is broadcast across the three image channels. The resulting image retains the dominant salient and near-field structure while suppressing much of the surrounding background, thereby reducing the visual ambiguity presented to the downstream zero-shot segmentation model.

3.3 Integration with Zero-Shot Segmentation Frameworks

The proposed conditioning module is independent of the downstream segmentation architecture. We integrate it with Grounded-SAM and SAM 3 to evaluate whether reducing background ambiguity benefits zero-shot systems with different localization and segmentation mechanisms.

3.3.1 SD-Grounded-SAM

SD-Grounded-SAM is a domain-aware extension of Grounded-SAM. In this configuration, Grounding DINO operates on the conditioned image $\mathbf{I}_{\mathrm{ref}}$ to generate candidate boxes for tower-mounted components. The candidate detections are subsequently refined using tower-relative geometry, pairwise containment, and relative-depth cues. The refined boxes are then provided to SAM as spatial prompts for instance-mask generation on the original image.

Open-Vocabulary Detection for Box-Prompt Generation. To automatically generate bounding-box prompts, we employ Grounding DINO [13], an open-vocabulary object detector that localizes objects from natural-language queries. The conditioned foreground image $\mathbf{I}_{\mathrm{ref}}$ is provided to Grounding DINO together with the text prompt $\mathbf{p}$ describing the target tower-mounted components. The detector produces a set of candidate bounding boxes

$$\mathcal{B} = \{\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_N\}. \tag{12}$$

Although foreground conditioning suppresses much of the irrelevant scene content, the detector may still produce false-positive, redundant, or overly broad candidate boxes. The detections are therefore passed to a subsequent box-refinement stage before being used as spatial prompts for SAM.

Bounding-Box Post-processing. To improve the quality of the spatial prompts supplied to SAM, we apply a structured post-processing procedure based on box size, pairwise containment, and relative depth. First, we remove detections that are disproportionately large relative to the estimated tower region. To define this reference region, we compute the tight bounding box enclosing the final refined foreground mask $\mathbf{M}_{\mathrm{ref}}$. Because individual tower-mounted components are expected to be substantially smaller than the tower structure itself, candidate boxes whose width or height exceeds a fraction $\tau_s \in [0, 1]$ of the corresponding tower bounding-box dimension are discarded. These detections typically correspond to the entire tower or to large structural regions rather than individual antennas or radio units.

Next, we remove redundant or nested detections using an asymmetric pairwise containment measure. For two boxes $\mathbf{B}_i$ and $\mathbf{B}_j$, we define

$$C(\mathbf{B}_i, \mathbf{B}_j) = \frac{|\mathbf{B}_i \cap \mathbf{B}_j|}{|\mathbf{B}_i|}, \tag{13}$$

where $|\cdot|$ denotes box area. Unlike standard Intersection over Union (IoU), this measure quantifies the proportion of $\mathbf{B}_i$ contained within $\mathbf{B}_j$ and is therefore suitable for identifying nested detections of different sizes. If

$$C(\mathbf{B}_i, \mathbf{B}_j) > \tau_{\text{iou}}, \tag{14}$$

the lower-confidence or otherwise redundant box is removed, where $\tau_{\text{iou}} \in [0, 1]$ is the containment threshold.

Finally, we use the relative-depth map to suppress detections located farther from the camera. For each remaining box $\mathbf{B}_i$, the mean depth value is computed over the pixels inside the box:

$$\mu_i = \frac{1}{|\Omega_i|} \sum_{(x,y)\in\Omega_i} \mathbf{D}(x, y), \tag{15}$$

where Ω_i denotes the set of pixels enclosed by $\mathbf{B}_i$. Boxes whose mean depth is classified as background according to the threshold $\tau_d \in [0, 255]$ are discarded.

This depth-based filtering is particularly useful in images captured from elevated or near top-down viewpoints, where rooftops, ground objects, and other background structures may exhibit shapes similar to tower-mounted components and may remain after foreground conditioning. The complete post-processing operation produces the refined box set

$$\mathcal{B}_{\text{ref}} = \mathcal{P}\left(\mathcal{B}, \mathbf{D}; \tau_s, \tau_{\text{iou}}, \tau_d\right), \tag{16}$$

where $\mathcal{P}(\cdot)$ denotes the box-refinement procedure.

Prompt-Based Segmentation. We use the SAM [12] to generate instance-level masks from the refined bounding boxes. Each refined bounding box $\mathbf{B}_i \in \mathcal{B}_{\text{ref}}$ is provided to SAM as a spatial prompt together with the original RGB image $\mathbf{I}$:

$$\mathbf{M}_i^{\text{GSAM}} = \text{SAM}\left(\mathbf{I}, \mathbf{B}_i\right), \tag{17}$$

where $\mathbf{M}_i^{\text{GSAM}} \in \{0, 1\}^{H\times W}$ denotes the predicted binary mask for the ith retained component. Applying SAM to all refined boxes produces the final SD-Grounded-SAM output:

$$\mathcal{M}_{\text{GSAM}} = \left\{\mathbf{M}_1^{\text{GSAM}}, \mathbf{M}_2^{\text{GSAM}}, \ldots, \mathbf{M}_{K_{\text{G}}}^{\text{GSAM}}\right\}, \tag{18}$$

where $K_{\text{G}} = |\mathcal{B}_{\text{ref}}|$ is the number of retained detections after post-processing. Although box generation is performed on the conditioned image, mask prediction is applied to the original RGB image to preserve the full visual detail of each component.

3.3.2 SD-SAM 3

We also integrate the same saliency–depth conditioning module with SAM 3, resulting in SD-SAM 3. Unlike the Grounded-SAM-based pipeline, SAM 3 performs concept localization and instance segmentation internally and therefore does not require externally generated bounding-box prompts. The conditioned foreground image $\mathbf{I}_{\text{ref}}$ and the same textual prompt $\mathbf{p}$ are supplied directly to the SAM 3 concept-prompted inference pipeline:

$$\mathcal{M}_{\text{SAM3}} = \text{SAM3}\left(\mathbf{I}_{\text{ref}}, \mathbf{p}\right), \tag{19}$$

where

$$\mathcal{M}_{\text{SAM3}} = \left\{\mathbf{M}_1^{\text{SAM3}}, \mathbf{M}_2^{\text{SAM3}}, \ldots, \mathbf{M}_{K_{\text{S}}}^{\text{SAM3}}\right\} \tag{20}$$

denotes the set of instance masks generated by SAM 3. No external box-generation or box-refinement stage is applied in this branch. Instead, the conditioning module modifies the visual input presented to SAM 3, while concept localization, instance identification, and mask generation remain internal to the model. This integration allows us to isolate the effect of saliency–depth foreground conditioning on a unified zero-shot segmentation architecture and to assess whether its benefit extends beyond the Grounded-SAM pipeline.

(a) (b)

Figure 2: **Representative instance-level annotations from TOW-300.** All component instances are shown using the same overlay color (pink) for improved visibility and visual clarity.

3.4 Model Modularity

Although the experiments in this work use Transparent Background for saliency estimation, Depth Anything for monocular depth prediction, Grounding DINO for open-vocabulary localization, and SAM/SAM 3 for segmentation, the proposed framework is not tied to these specific models. Each component can be replaced by an alternative model that produces a compatible intermediate output, such as a saliency map, relative-depth map, candidate bounding boxes, or instance masks. The main contribution therefore lies in the saliency–depth conditioning and integration strategy rather than in any particular model choice. The use of two distinct downstream segmentation frameworks further demonstrates this modularity.

4 Dataset

We construct TOW-300, a UAV-based dataset for instance-level segmentation of communication-tower components. The dataset is collected from three separate UAV flights covering two visually and structurally distinct communication towers and contains 340 high-resolution images with multiple mounted components, primarily antennas and radio units. Each visible component is treated as an individual instance. The images have resolutions of either 5000×6300 or 9100×6300 pixels, preserving the fine-grained visual details required to delineate small and partially occluded components.

Data Collection. The images were collected during real-world tower inspections using a DJI Mavic 3 Enterprise UAV equipped with a 20 MP camera. [41]. To capture variation representative of practical inspection conditions, the images were acquired at different times of day and under varying illumination conditions. The UAV followed multiple flight trajectories, including helical paths around the tower and vertical lawnmower patterns, producing observations from different viewpoints, elevations, and distances. Consequently, the dataset exhibits substantial variation in component scale, perspective, occlusion, and background composition.

Annotation. All images were annotated using the CVAT platform [42]. Instance-level polygon masks were manually drawn around each visible tower-mounted component and exported in COCO format, including both segmentation masks and their corresponding bounding boxes [43]. The dataset contains a single semantic class, *tower-mounted component*, which includes antennas and radio units, while preserving a separate annotation for each component instance. Partially visible components were annotated when their boundaries could be identified with sufficient confidence. Components for which only a small and highly ambiguous portion was visible were excluded to avoid introducing unreliable annotations. This policy was applied consistently throughout the dataset. Representative ground-truth annotations are shown in Figure 2.

Dataset Characteristics and Challenges. TOW-300 contains several characteristics, making instance segmentation challenging, as illustrated in Figure 3. Communication towers commonly have sparse lattice or truss structures through which vegetation, buildings, rooftops, cables, and other background objects remain visible. These background elements may exhibit shapes, textures, and colors similar to those of tower-mounted components, increasing the ambiguity between foreground components and surrounding structures. The dataset also includes substantial variation in viewpoint,

(a) (b) (c) (d)

Figure 3: **Representative challenges in TOW-300.** (a) Background structures remain visible through the sparse tower lattice and exhibit colors and appearances similar to tower-mounted components. (b) A small background structure resembles the communication tower, creating ambiguity for coarse text prompts such as *antenna*. (c) The tower occupies only a small portion of the full high-resolution image and is surrounded by substantial background clutter. (d) A near top-down view exposes rooftop objects through the sparse tower structure, some of which visually resemble mounted communication equipment.

scale, and perspective, including elevated and near top-down observations. In some images, the tower occupies only a small portion of the full scene, while in others, background structures resembling tower elements remain clearly visible through or around the tower. Individual components may therefore appear small, overlap with structural members, or be partially occluded by neighboring equipment, further complicating reliable instance localization and segmentation.

These challenges are particularly pronounced in a zero-shot setting. Open-vocabulary detectors rely on general-purpose textual descriptions and may not possess sufficient domain knowledge to distinguish individual antennas and radio units from the tower structure or visually similar background objects. Consequently, generic or appearance-based prompts may produce false-positive detections on rooftops, windows, signs, vegetation, and metallic structures. TOW-300 therefore provides a challenging real-world setting for evaluating zero-shot component localization and segmentation under substantial visual clutter.

Data Split. Because our proposed framework does not require model training, the dataset is divided into validation and test subsets rather than conventional training, validation, and test sets. We use 40 images to select the inference-time hyperparameters of the proposed pipeline, including the saliency–depth recovery, box-size, containment, and depth-filtering thresholds. The remaining 300 images are reserved for final evaluation. All hyperparameters are selected using only the validation subset and are subsequently fixed for evaluation on the test set.

5 Experiments

5.1 Experimental Setup

Compared Configurations. We evaluate the proposed saliency–depth conditioning strategy with two existing zero-shot segmentation frameworks: Grounded-SAM and SAM 3. For each framework, we compare the original model with its conditioned counterpart, resulting in four evaluated methods: Grounded-SAM, SD-Grounded-SAM, SAM 3, and SD-SAM 3. This paired comparison isolates the effect of saliency–depth conditioning within each framework. SD-Grounded-SAM additionally includes the proposed geometric and depth-based box-refinement stage before mask generation, whereas SD-SAM 3 relies on SAM 3's internal localization and segmentation mechanism without external box refinement.

Evaluation Protocol. The proposed methods operate in a training-free setting, and no model parameters are updated using TOW-300. All models are evaluated in inference mode using publicly available pre-trained weights. The dataset is divided into a validation subset of $N_{\text{val}} = 40$ images for inference-time hyperparameter selection and a test subset of $N_{\text{test}} = 300$ images for final evaluation. All hyperparameters are selected exclusively on the validation subset and fixed before evaluation on the test set. The same validation and test subsets are used for all compared methods.

Implementation Details. All experiments are conducted on the original high-resolution images without resizing, thereby preserving the fine-grained details of small tower-mounted components.

Inference is performed on a single NVIDIA RTX 4090 GPU, and images are processed individually without batching. We use publicly available implementations and pre-trained checkpoints for the Transparent Background saliency model [15], Depth Anything [16], Grounding DINO [13], SAM [12], and SAM 3 [38].

Prompt and Model Configuration. We use the fixed textual prompt *"small bright rectangles attached to tower"* for all compared methods. The prompt describes the visual appearance and placement of the target tower-mounted components and is kept unchanged across methods to avoid introducing differences caused by prompt selection.

For Grounded-SAM and SD-Grounded-SAM, Grounding DINO with a Swin-T backbone is used for open-vocabulary detection, followed by SAM ViT-H for mask generation, whereas SAM 3 and SD-SAM 3 are evaluated through SAM 3's zero-shot concept-prompted segmentation pipeline. The conditioned variants share the same foreground conditioning configuration, using Depth Anything ViT-L for monocular depth estimation and the base Transparent Background model for saliency prediction.

Hyperparameter Selection. The saliency–depth conditioning module and the SD-Grounded-SAM box-refinement stage involve several inference-time hyperparameters introduced in the Method section. These include the depth-recovery threshold τ_{rec}, the relative box-size threshold τ_s, the containment threshold τ_{iou}, and the box-level depth threshold τ_d. The values selected on the validation subset are

$$\tau_{\mathrm{rec}} = 140, \qquad \tau_s = 0.4, \qquad \tau_{\mathrm{iou}} = 0.5, \qquad \tau_d = 70. \tag{21}$$

These values are selected using only the validation subset and remain fixed during final testing. Each segmentation framework also uses a model-specific confidence threshold to determine which predicted instances are retained. We select this threshold through a grid search on the validation subset, using the value that maximizes the F1-score at an instance-mask IoU threshold of 0.5. The same selection procedure is applied to the original and conditioned variants of each framework. The resulting confidence thresholds are 0.14 for Grounded-SAM and SD-Grounded-SAM, and 0.20 for SAM 3 and SD-SAM 3. These thresholds are also fixed during final evaluation.

5.2 Evaluation Metrics

We evaluate performance using both instance-level and semantic segmentation metrics to provide a comprehensive assessment of detection and mask quality.

Instance Segmentation Metrics. We report standard instance segmentation metrics, including Average Precision (AP) at different IoU thresholds. Specifically, we use AP@0.5, AP@0.75, and the COCO-style mean Average Precision (mAP@[0.5:0.95]) [43]. These metrics evaluate the quality of both localization and segmentation across varying levels of overlap between predicted and ground-truth instances. In addition, we report Precision@0.5 and Recall@0.5 to analyze detection performance at IoU=0.5. To further assess mask quality for correctly matched instances, we compute the mean matched IoU, which measures the average overlap between predicted and ground-truth masks for true positive detections.

Semantic Segmentation Metrics. To evaluate overall mask quality at the image level, we also compute semantic segmentation metrics by merging all predicted instance masks into a single binary mask. We report IoU and Dice coefficient for the *tower-mounted components* class. These metrics capture the overall coverage and accuracy of the predicted foreground regions.

5.3 Quantitative Results

Tables 1 and 2 report the instance-level and image-level segmentation results, respectively. The proposed saliency–depth conditioning strategy improves both evaluated zero-shot segmentation frameworks across all reported metrics. SD-SAM 3 achieves the strongest overall instance segmentation performance, obtaining the highest mAP, AP@0.75, AP@0.5, recall, and mean matched IoU. In contrast, SD-Grounded-SAM achieves the highest precision and the strongest image-level semantic segmentation results. These findings show that foreground conditioning provides consistent benefits across architecturally distinct segmentation systems, while the downstream framework and box-refinement design influence the resulting precision–recall characteristics.

Table 1: Instance segmentation performance on the TOW-300 test set. Mean matched IoU is computed over true-positive instance matches at an IoU threshold of 0.5. Best results are shown in bold, and second-best results are underlined.

Method	mAP@[0.5:0.95] (↑)	AP@0.75 (↑)	@IoU=0.5			
			AP (↑)	Recall (↑)	Precision (↑)	Matched IoU (↑)
Grounded-SAM	0.1737	0.1819	0.2236	0.3971	0.4962	0.8879
SAM 3	<u>0.4250</u>	<u>0.4831</u>	<u>0.5438</u>	<u>0.6089</u>	0.7662	0.8787
SD-Grounded-SAM (Ours)	0.3780	0.4129	0.4751	0.5073	**0.8970**	<u>0.8989</u>
SD-SAM 3 (Ours)	**0.5701**	**0.6270**	**0.6709**	**0.7311**	<u>0.7817</u>	**0.9107**

Table 2: Semantic segmentation results obtained by merging all predicted instance masks into a binary foreground mask. Best results are shown in bold, and second-best results are underlined.

Method	IoU (↑)	Dice (↑)
Grounded-SAM	0.2876	0.4467
SAM 3	0.3630	0.5326
SD-Grounded-SAM (Ours)	**0.6281**	**0.7716**
SD-SAM 3 (Ours)	<u>0.4376</u>	<u>0.6088</u>

Effect on Grounded-SAM. SD-Grounded-SAM substantially improves its direct Grounded-SAM baseline across every instance-level metric. COCO-style mAP increases from 0.1737 to 0.3780, while AP@0.5 increases from 0.2236 to 0.4751 and AP@0.75 increases from 0.1819 to 0.4129. These correspond to relative improvements of approximately 117.6%, 112.5%, and 127.0%, respectively. Recall also increases from 0.3971 to 0.5073, indicating that the conditioned pipeline recovers more valid component instances than the original Grounded-SAM configuration.

The largest improvement is observed in precision, which increases from 0.4962 to 0.8970. This result indicates that the combined foreground-conditioning and box-refinement stages effectively suppress false detections arising from visually similar regions. Mean matched IoU also increases from 0.8879 to 0.8989. Although this gain is smaller, it shows that the refined box prompts preserve, and slightly improve, the quality of masks associated with correctly localized instances. Overall, the results suggest that the main weakness of Grounded-SAM in this domain lies in candidate localization and prompt generation rather than in the mask-generation capability of SAM itself.

Effect on SAM 3. Applying the same saliency–depth conditioning strategy to SAM 3 also produces consistent improvements, despite the model performing localization and segmentation internally and receiving no external box refinement. SD-SAM 3 increases mAP from 0.4250 to 0.5701, AP@0.5 from 0.5438 to 0.6709, and AP@0.75 from 0.4831 to 0.6270. These correspond to relative gains of approximately 34.1%, 23.4%, and 29.8%, respectively. Recall increases from 0.6089 to 0.7311, showing that conditioning allows SAM 3 to identify a substantially larger proportion of the annotated components.

Precision also improves from 0.7662 to 0.7817, while mean matched IoU increases from 0.8787 to 0.9107. The simultaneous gains in precision and recall indicate that foreground conditioning improves the quality of concept localization rather than simply making the model more conservative or more permissive. By suppressing irrelevant visual content while preserving tower-relevant regions, the conditioned input helps SAM 3 identify more valid components and generate more accurate masks. Since SD-SAM 3 does not use the external box-refinement stage of SD-Grounded-SAM, these improvements can be attributed directly to the saliency–depth conditioning module.

Comparison of the Conditioned Frameworks. Among the conditioned methods, SD-SAM 3 achieves higher instance-level AP and recall than SD-Grounded-SAM. This indicates that SAM 3's integrated concept-localization mechanism is more effective at discovering small, partially occluded, or visually weak component instances once background interference has been reduced.

SD-Grounded-SAM, however, achieves the highest precision. This difference is consistent with its explicit box-refinement stage, which removes detections based on tower-relative size, pairwise containment, and relative depth before mask generation. Consequently, SD-Grounded-SAM adopts a more conservative operating profile, retaining fewer predictions but producing a larger proportion

of valid component detections. The two conditioned configurations are therefore complementary: SD-SAM 3 is preferable when component discovery and recall are prioritized, whereas SD-Grounded-SAM is advantageous when false-positive suppression is more important.

Image-Level Foreground Quality. The semantic results in Table 2 further demonstrate the effect of the proposed conditioning strategy. SD-Grounded-SAM improves semantic IoU from 0.2876 to 0.6281 and Dice from 0.4467 to 0.7716 relative to Grounded-SAM. These correspond to relative improvements of approximately 118.4% and 72.7%, respectively. The large gains indicate that foreground conditioning and explicit box refinement substantially reduce the total area assigned to irrelevant background objects.

SD-SAM 3 also improves consistently over SAM 3, increasing semantic IoU from 0.3630 to 0.4376 and Dice from 0.5326 to 0.6088. These gains of approximately 20.6% and 14.3% show that conditioning produces cleaner aggregate masks even when object localization remains internal to SAM 3. However, SD-Grounded-SAM achieves the strongest semantic results overall.

This difference between the instance-level and semantic results can be explained by the different prediction behavior of the two pipelines. SD-Grounded-SAM is deliberately conservative: its box-refinement stage removes many uncertain detections, including oversized, nested, and distant boxes, and therefore retains a smaller set of high-confidence component predictions. As a result, it may miss some valid instances, but it also produces substantially fewer false-positive masks, which leads to cleaner merged foreground predictions and stronger semantic IoU and Dice. In contrast, SD-SAM 3 attempts to recover a larger number of component instances and therefore achieves stronger instance-level performance, but this higher recall can also introduce additional false-positive or overly broad masks. In some cases, SD-SAM 3 predicts a large mask covering much of the tower in addition to the desired component-level masks, which increases the predicted foreground area and degrades image-level semantic metrics. A representative example of this behavior is shown in Fig. 4.

Discussion. Taken together, the results support two main conclusions. First, saliency–depth foreground conditioning addresses a general background-ambiguity problem rather than a limitation specific to a single segmentation architecture. Both Grounded-SAM and SAM 3 improve across instance- and image-level metrics when supplied with the conditioned image. Second, the downstream segmentation framework determines how these benefits are expressed. SD-SAM 3 provides the strongest overall instance discovery and mask overlap, while SD-Grounded-SAM combines conditioning with explicit box refinement to achieve the highest precision and cleanest aggregate foreground segmentation. These complementary behaviors provide different operating points for practical inspection workflows, depending on whether maximizing component coverage or minimizing false detections is the primary objective.

5.4 Qualitative Results

Figures 4 and 5 provide qualitative comparisons of the original and conditioned segmentation frameworks. Different colors are used to distinguish individual instances belonging to the same component class.

Effect of Saliency–Depth Conditioning. Figure 4 shows that saliency–depth conditioning improves both Grounded-SAM and SAM 3. The original Grounded-SAM frequently produces broad masks covering large portions of the tower and does not accurately separate individual components. SD-Grounded-SAM substantially reduces these oversized predictions and produces more compact masks that better match the annotated component regions. Similarly, SAM 3 generally localizes more instances than Grounded-SAM, but it may also respond to visually similar background objects, as observed in the second and fourth rows. SD-SAM 3 reduces much of this background confusion while preserving SAM 3's strong instance-localization capability. These visual improvements are consistent with the quantitative results, where both conditioned variants outperform their corresponding unconditioned baselines.

Failure Cases. Figure 5 presents two representative challenging cases. In the first and second rows, SD-SAM 3 correctly localizes several component instances but also produces an oversized mask covering a substantial portion of the tower. When all predicted instances are merged into a binary foreground mask, this prediction introduces considerable non-target area and reduces image-level IoU and Dice. In contrast, SD-Grounded-SAM suppresses the tower-scale detection through its

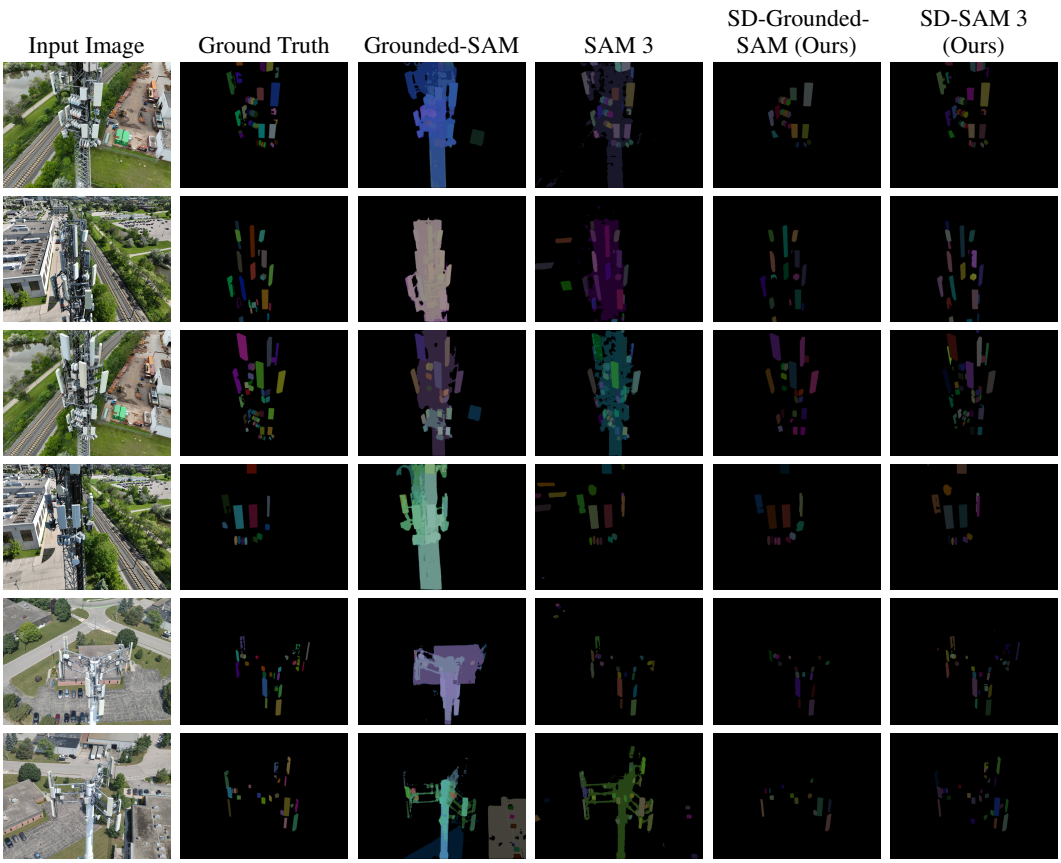

Figure 4: Qualitative comparison of Grounded-SAM, SAM 3, and their saliency–depth-conditioned variants. Each row presents a representative UAV image, its ground-truth annotation, and the corresponding predictions.

box-refinement stage and produces a cleaner result. This example supports the quantitative finding that SD-SAM 3 achieves higher recall and instance-level AP, whereas SD-Grounded-SAM achieves higher precision and semantic foreground quality.

The third and fourth rows illustrate the difficulty of near top-down viewpoints. From this perspective, the tower overlaps visually with rooftops and other surrounding structures, making it difficult for saliency to separate the foreground accurately. As a result, some background content can remain in the conditioned image and influence the final predictions. SD-SAM 3 recovers several plausible component instances but also includes predictions associated with the retained background. SD-Grounded-SAM again produces a more selective result because its geometric and depth-based box refinement removes many unreliable detections, although this conservative behavior may also cause some valid components to be missed.

Overall, the qualitative comparisons confirm that saliency–depth conditioning reduces background ambiguity in both frameworks, while the downstream localization and refinement mechanisms determine the final balance between instance coverage and false-positive suppression.

5.5 Ablation Study

Table 3 evaluates the individual contributions of saliency conditioning, depth-based recovery, and box refinement across Grounded-SAM and SAM 3. The shared conditioning components are progressively introduced in both frameworks, while paired Grounded-SAM configurations with and without box refinement isolate the effect of the additional prompt-filtering stage. We report mAP, precision, recall, and semantic IoU to capture instance discovery, false-positive suppression, and aggregate foreground quality.

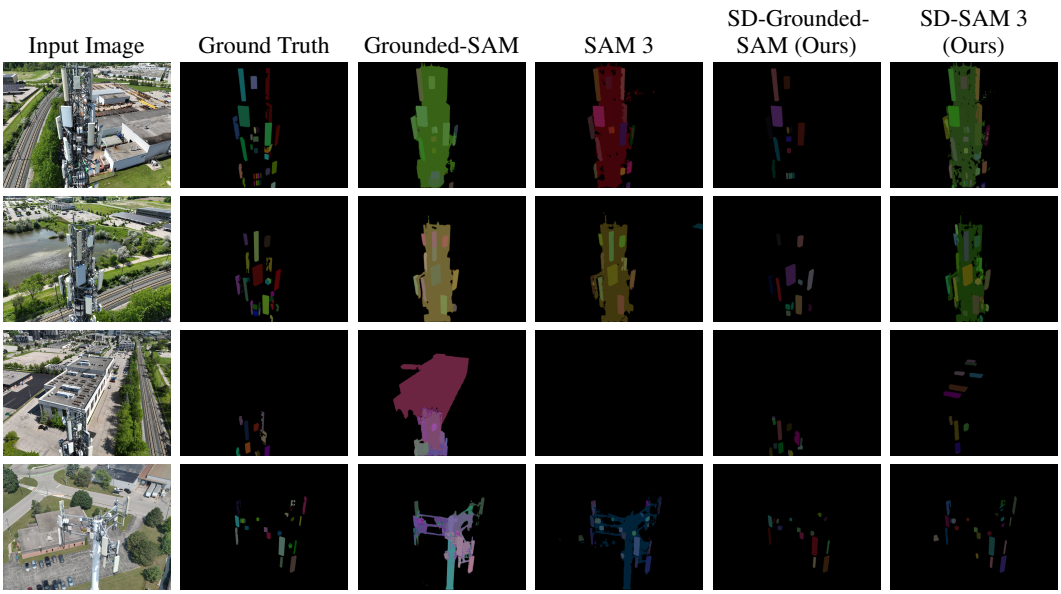

Figure 5: **Representative failure cases.** In the first and second rows, SD-SAM 3 detects several valid components but also produces an oversized tower-level mask, whereas SD-Grounded-SAM suppresses this prediction through box refinement. The third and fourth rows show near top-down views in which foreground conditioning retains visually similar background regions, leading to localization errors.

Table 3: Ablation study of saliency–depth conditioning and box refinement across the Grounded-SAM and SAM 3 frameworks. Best and second-best results are identified separately within each framework and are shown in bold and underlined, respectively.

Model	Conditioning	Box Refinement	mAP (↑)	Precision@0.5 (↑)	Recall@0.5 (↑)	Semantic IoU (↑)
Grounded-SAM	None	✓	0.2870	0.8787	0.3920	0.6195
Grounded-SAM	Saliency	✓	<u>0.3355</u>	<u>0.8845</u>	0.4511	**0.6420**
Grounded-SAM	Saliency + depth	✗	0.2569	0.5499	**0.5226**	0.2920
Grounded-SAM	Saliency + depth	✓	**0.3780**	**0.8970**	<u>0.5073</u>	<u>0.6281</u>
SAM 3	None	–	0.4250	0.7662	0.6089	0.3630
SAM 3	Saliency	–	<u>0.5681</u>	<u>0.7815</u>	<u>0.7281</u>	**0.4420**
SAM 3	Saliency + depth	–	**0.5701**	**0.7817**	**0.7311**	<u>0.4376</u>

Effect of Foreground Conditioning. For Grounded-SAM, saliency increases mAP from 0.2870 to 0.3355 and recall from 0.3920 to 0.4511 when box refinement is enabled. Adding depth further increases mAP to 0.3780 and recall to 0.5073. A similar trend is observed for SAM 3, where saliency increases mAP from 0.4250 to 0.5681 and recall from 0.6089 to 0.7281, while adding depth further improves them to 0.5701 and 0.7311, respectively. These results indicate that saliency reduces background ambiguity, whereas depth helps recover additional near-field component regions that may be missed by saliency alone. The small decrease in semantic IoU after adding depth suggests that this higher instance coverage can also retain some nearby non-target regions.

Effect of Box Refinement. The importance of box refinement is evident when comparing the two saliency–depth-conditioned Grounded-SAM configurations. Without box refinement, recall reaches 0.5226, slightly higher than the 0.5073 obtained by the complete pipeline. However, precision decreases sharply from 0.8970 to 0.5499, mAP falls from 0.3780 to 0.2569, and semantic IoU drops from 0.6281 to 0.2920. This shows that the refinement stage effectively removes oversized, redundant, and background detections, producing a cleaner set of spatial prompts with only a small loss in recall.

Overall Findings. Overall, the ablation study shows that saliency is the primary source of improvement across both segmentation frameworks, while depth complements it by recovering additional near-field regions and improving instance discovery. Box refinement suppresses unreliable detections, thereby substantially increasing precision and aggregate foreground quality. Together, these compo-

nents provide complementary benefits: saliency–depth conditioning improves component discovery across both frameworks, while the Grounded-SAM-specific refinement stage improves the reliability of the final predictions.

6 Conclusion

This work introduced a model-agnostic saliency–depth foreground-conditioning strategy for training-free segmentation of communication-tower components in cluttered UAV imagery. The proposed conditioning module combines appearance-based saliency with monocular relative depth to suppress irrelevant background regions while recovering near-field tower structures that may be missed by saliency alone. We integrated this strategy with two architecturally distinct zero-shot segmentation frameworks, resulting in SD-Grounded-SAM and SD-SAM 3. For the Grounded-SAM branch, we additionally introduced a geometric and depth-aware box-refinement stage to improve the reliability of the spatial prompts provided to SAM. Experiments on the author-constructed TOW-300 dataset demonstrate that the proposed conditioning strategy consistently improves both underlying frameworks across instance- and image-level segmentation metrics. SD-SAM 3 achieves the strongest overall instance segmentation performance, showing that foreground conditioning improves SAM 3's internal concept localization and component discovery. SD-Grounded-SAM achieves the highest precision and semantic IoU, reflecting the effectiveness of its explicit box-refinement stage in suppressing oversized, redundant, and background detections. The ablation study further shows that saliency reduces background ambiguity, depth improves instance discovery by recovering additional near-field regions, and box refinement substantially improves false-positive suppression and aggregate foreground quality. These findings indicate that background conditioning addresses a general limitation of zero-shot segmentation in visually complex infrastructure scenes rather than a weakness specific to a single model. Nevertheless, depth-based recovery may retain some nearby non-target structures, while conservative box filtering can remove small or weakly visible components. Future work will investigate adaptive fusion and filtering strategies, improved foreground-prior estimation, and evaluation across additional tower configurations and infrastructure datasets to further improve the balance between component discovery and false-positive suppression.

References

[1] C. Lyu, S. Lin, A. Lynch, Y. Zou, and M. Liarokapis, "Uav-based deep learning applications for automated inspection of civil infrastructure," *Automation in Construction*, vol. 177, p. 106285, 2025.

[2] M. A. A. Faisal, I. Mecheter, Y. Qiblawey, J. H. Fernandez, M. E. Chowdhury, and S. Kiranyaz, "Deep learning in automated power line inspection: A review," *Applied Energy*, vol. 385, p. 125507, 2025.

[3] A. Villarino, H. Valenzuela, N. Antón, M. Domínguez, and X. C. Méndez Cubillos, "Uav applications for monitoring and management of civil infrastructures," *Infrastructures*, vol. 10, no. 5, p. 106, 2025.

[4] X. Li, Y. Li, Y. Chen, G. Zhang, and Z. Liu, "Deep learning-based target point localization for uav inspection of point cloud transmission towers," *Remote Sensing*, vol. 16, no. 5, p. 817, 2024.

[5] H. Wang, Q. Yang, B. Zhang, and D. Gao, "Deep learning based insulator fault detection algorithm for power transmission lines," *Journal of Real-Time Image Processing*, vol. 21, no. 4, p. 115, 2024.

[6] M. Liu, Y. Li, X. Wang, R. Tu, and Z. Zhu, "Deep learning-based segmentation of key objects of transmission lines," in *International Conference on Entertainment Computing*, pp. 317–324, Springer, 2020.

[7] R. R. Nejadbougar, E. G. Parmehr, A. Afary, and S. Mavaddati, "Intelligent monitoring of power transmission line insulators using uav images and deep learning model,"

[8] Q. Wang, Y. Li, S. Cui, N. Li, X. Zhang, W. Jiang, W. Peng, and J. Sun, "Enhanced recognition of insulator defects on power transmission lines via proposal-based detection model with integrated improvement methods," *Engineering Applications of Artificial Intelligence*, vol. 136, p. 109078, 2024.

[9] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9157–9166, 2019.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[11] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," *Advances in Neural information processing systems*, vol. 33, pp. 17721–17732, 2020.

[12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

[13] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European conference on computer vision*, pp. 38–55, Springer, 2024.

[14] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded sam: Assembling open-world models for diverse visual tasks," 2024.

[15] T. Kim, K. Kim, J. Lee, D. Cha, J. Lee, and D. Kim, "Revisiting image pyramid structure for high resolution salient object detection," in *Proceedings of the Asian Conference on Computer Vision*, pp. 108–124, 2022.

[16] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10371–10381, 2024.

[17] K. Messaoudi, O. S. Oubbati, A. Rachedi, A. Lakas, T. Bendouma, and N. Chaib, "A survey of uav-based data collection: Challenges, solutions and future perspectives," *Journal of network and computer applications*, vol. 216, p. 103670, 2023.

[18] O. Bouhamed, H. Ghazzai, H. Besbes, and Y. Massoud, "A uav-assisted data collection for wireless sensor networks: Autonomous navigation and scheduling," *IEEe Access*, vol. 8, pp. 110446–110460, 2020.

[19] S. Alfattani, W. Jaafar, H. Yanikomeroglu, and A. Yongacoglu, "Multi-uav data collection framework for wireless sensor networks," in *2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2019.

[20] R. Pomarnacki, D. Bručas, T. Jačionis, and D. Plonis, "Unmanned aerial vehicles for the monitoring of telecommunication towers from the engineering approach," *Drones*, vol. 9, no. 4, p. 308, 2025.

[21] M. Korki, N. D. Shankar, R. N. Shah, S. M. Waseem, and S. Hodges, "Automatic fault detection of power lines using unmanned aerial vehicle (uav)," in *2019 1st International Conference on Unmanned Vehicle Systems-Oman (UVS)*, pp. 1–6, IEEE, 2019.

[22] S. Farhangfar and M. Rezaeian, "Semantic segmentation of aerial images using fcn-based network," in *2019 27th Iranian conference on electrical engineering (ICEE)*, pp. 1864–1868, IEEE, 2019.

[23] I. O. Agyemang, X. Zhang, I. Adjei-Mensah, D. Acheampong, L. D. Fiasam, C. Sey, S. B. Yussif, and D. Effah, "Automated vision-based structural health inspection and assessment for post-construction civil infrastructure," *Automation in Construction*, vol. 156, p. 105153, 2023.

[24] J.-L. Xiao, J.-S. Fan, Y.-F. Liu, B.-L. Li, and J.-G. Nie, "Region of interest (roi) extraction and crack detection for uav-based bridge inspection using point cloud segmentation and 3d-to-2d projection," *Automation in Construction*, vol. 158, p. 105226, 2024.

[25] A. K. Sangaiah, J. Anandakrishnan, N. K. Son, H. Darmawan, G.-B. Bian, and M. J. Alenazi, "Lcut-sv9: Uav-assisted powerline inspection framework with secure time-sensitive communication for industry 5.0," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 2837–2852, 2025.

[26] K. Lan, X. Zheng, S. Li, J. Hu, and L. Shao, "Amfa-net: Aerial power line segmentation network based on attention mechanism and feature aggregation," *Measurement*, p. 118684, 2025.

[27] Q. Zhao, H. Fang, Y. Pang, G. Zhu, and Z. Qian, "Pl-unet: a real-time power line segmentation model for aerial images based on adaptive fusion and cross-stage multi-scale analysis," *Journal of Real-Time Image Processing*, vol. 22, no. 1, p. 39, 2025.

[28] R. Abdelfattah, X. Wang, and S. Wang, "Ttpla: An aerial-image dataset for detection and segmentation of transmission towers and power lines," in *Proceedings of the Asian conference on computer vision*, 2020.

[29] Y. Si, J. Gao, M. Zhao, and X. Xu, "Research on the algorithm of detecting insulators in high-voltage transmission lines using uav images," *Signal, Image and Video Processing*, vol. 18, no. Suppl 1, pp. 395–406, 2024.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[31] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[32] M. Calabrese, L. Agnusdei, G. Fontana, G. Papadia, and A. Del Prete, "Application of mask r-cnn and yolov8 algorithms for defect detection in printed circuit board manufacturing," *Discover Applied Sciences*, vol. 7, no. 4, p. 257, 2025.

[33] X. Xu, M. Zhao, P. Shi, R. Ren, X. He, X. Wei, and H. Yang, "Crack detection and comparison study based on faster r-cnn and mask r-cnn," *Sensors*, vol. 22, no. 3, p. 1215, 2022.

[34] G. Zhu, W. Zhang, M. Wang, J. Wang, and X. Fang, "Corner guided instance segmentation network for power lines and transmission towers detection," *Expert Systems with Applications*, vol. 234, p. 121087, 2023.

[35] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7086–7096, 2022.

[36] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," *arXiv preprint arXiv:2201.03546*, 2022.

[37] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "Groupvit: Semantic segmentation emerges from text supervision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18134–18144, 2022.

[38] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, *et al.*, "Sam 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.

[39] W. Zhou, H. Huang, H. Zhang, and C. Wang, "Teaching segment-anything-model domain-specific knowledge for road crack segmentation from on-board cameras," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 12, pp. 20588–20601, 2024.

[40] X. Pu, H. Jia, L. Zheng, F. Wang, and F. Xu, "Classwise-sam-adapter: Parameter-efficient fine-tuning adapts segment anything to sar domain for semantic segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 4791–4804, 2025.

[41] DJI, "DJI official website." `https://www.dji.com/ca`. Accessed: 2026-08-19.

[42] CVAT.ai, "CVAT: Computer vision annotation tool." `https://github.com/cvat-ai/cvat/`.

[43] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.