2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

# Vision-Language Feature Alignment for Road Anomaly Segmentation

Zhuolin He, Jiacheng Tang, Jian Pu*, Xiangyang Xue*

***Abstract*— Safe autonomous systems in complex environments require robust road anomaly segmentation to identify unknown obstacles. However, existing approaches often rely on pixel-level statistics to determine whether a region appears anomalous. This reliance leads to high false-positive rates on semantically normal background regions such as sky or vegetation, and poor recall of true Out-of-distribution (OOD) instances, thereby posing safety risks for robotic perception and decision-making. To address these challenges, we propose VL-Anomaly, a vision-language anomaly segmentation framework that incorporates semantic priors from pre-trained Vision-Language Models (VLMs). Specifically, we design a prompt learning-driven alignment module that adapts Mask2Former's visual features to CLIP text embeddings of known categories, effectively suppressing spurious anomaly responses in background regions. At inference time, we further introduce a multi-source inference strategy that integrates text-guided similarity, CLIP-based image-text similarity and detector confidence, enabling more reliable anomaly prediction by leveraging complementary information sources. Extensive experiments demonstrate that VL-Anomaly achieves state-of-the-art performance on benchmark datasets including RoadAnomaly, SMIYC and Fishyscapes. Code is released on `https://github.com/NickHezhuolin/VL-aligner-Road-anomaly-segment`.**

## I. INTRODUCTION

Semantic segmentation plays a pivotal role in enabling autonomous driving systems [1]–[3] and mobile robots [4] to achieve a fine-grained understanding of their surroundings. Typically, segmentation models are trained to recognize a fixed set of pre-defined categories [5], [6]. However, in real-world scenarios, these models often encounter out-of-distribution (OOD) objects, such as unexpected obstacles that were not present in the training data. Without proper mechanisms to identify such anomalies, models tend to misclassify these regions into known classes, resulting in inaccurate segmentation masks [7]. Such misclassification undermines model reliability and, more critically, poses substantial safety risks in safety-critical domains such as autonomous driving [8]. Given that fact, accurate segmentation of OOD anomalies is essential for building robust and trustworthy perception systems in open-world environments.

Road anomaly segmentation [9], [10] tackles this challenge by segmenting road objects that fall outside the set of known training classes. Most existing methods still follow a vision-only paradigm [11], [12], typically detecting anomalies by thresholding pixel-wise prediction confidence or measuring deviations in low-level visual features. Lacking high-level semantic understanding, these models are constrained to visual similarity, basing their predictions merely on resemblance to feature prototypes or distributions of known classes. Consequently, background regions such as sky and trees belong to normal classes in the dataset, but variations in texture or appearance such as clouds or colors can cause them to be falsely detected as anomalies, as demonstrated in Fig. 1. This phenomenon results in a high false-positive rate and undermines the robustness and practicality of existing approaches in real-world autonomous driving scenarios.

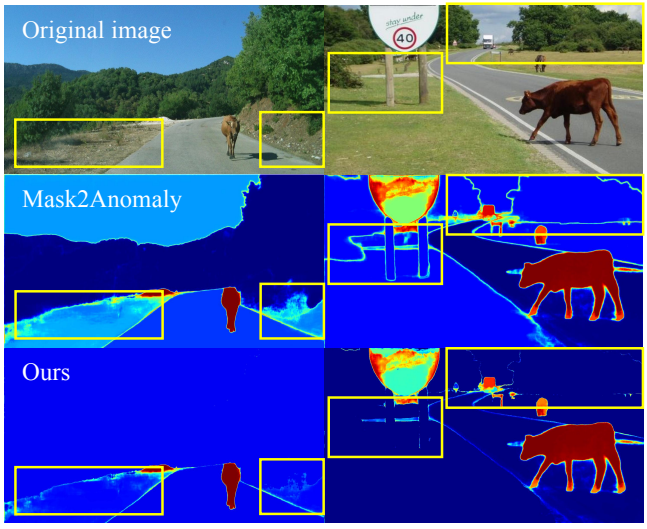


Fig. 1. Comparison of Anomaly Score Maps. The first row shows the original images, the second row presents anomaly score maps generated by Mask2Anomaly, and the third row illustrates the results of our method. Our approach yields cleaner maps by suppressing false positives on semantically normal background regions such as road surface and vegetation, while more precisely highlighting true anomalies like animals.

Recently, the emergence of vision-language models (VLMs) [13], [14] has introduced a new perspective to road anomaly segmentation. Textual guidance provides explicit semantics, allowing the model to identify pixels corresponding to known categories [15]. At the same time, the semantic dissimilarity derived from vision-language alignment naturally highlights regions that may correspond to unknown objects [16]. Leveraging semantic similarity not only suppresses false positives but also accentuates rare yet informative OOD cues, thereby enhancing detection accuracy and generalization in open-world scenarios.

Building on vision-language semantic cues, we propose VL-Anomaly, a vision-language anomaly segmentation framework that exploits VLM priors during both training and

Z. He, and X. Xue are with the School of Computer Science, Fudan University, Shanghai 200433, China. J. Tang and J. Pu are with ISTBI, Fudan University, Shanghai 200433, China. J. Pu is also with Embodiq Robotics Co., Ltd., Shanghai, China. This work is supported by NSFC General Program (Grant No. 62576110). * denotes the corresponding author. (E-mail: {jianpu, xyxue}@fudan.edu.cn).

inference. Its core idea is to integrate category-level knowledge from CLIP [13] into the segmentation process, thereby enhancing the model's ability to distinguish in-distribution categories from out-of-distribution regions. However, directly applying VLMs to segmentation is non-trivial, as they are not inherently designed for pixel-level multi-class prediction. The core challenge lies in adapting VLMs to the multi-class nature of semantic segmentation. Although recent approaches such as MaskCLIP [17] demonstrate that multiple categories can be matched within a single sentence, these methods typically rely on handcrafted or concatenated prompts that are not explicitly optimized for segmentation. To overcome this, we design the Prompt Learning-Driven Aligner (PL-Aligner). It employs learnable textual prompts to guide both the backbone features and decoder mask queries into the VLM semantic space. The alignment is carried out at two stages, pixel level and mask level, which together bridge the gap between visual and textual representations. PL-Aligner remains architecture-agnostic and can be seamlessly incorporated into existing segmentation frameworks without requiring structural modifications.

At inference time, we further enhance robustness by adopting a multi-source inference strategy. Specifically, we integrate (i) detector confidence from the segmentation network, (ii) text-guided similarity derived from learned prompts, and (iii) CLIP-based image–text similarity to exploit their complementary strengths. The fusion of these complementary signals provides reliable anomaly predictions and mitigates the weaknesses of relying on a single source. Through this multimodal integration, VL-Anomaly achieves fine-grained anomaly segmentation across diverse road anomaly datasets while maintaining strong generalization. To summarize, we make the following contributions:

- We propose PL-Aligner, a prompt-driven alignment module that jointly aligns features at both pixel and mask levels, leading to more robust text-guided anomaly segmentation.
- We introduce a multi-source inference strategy that integrates text-guided similarity, CLIP-based image-text similarity and detector confidence to deliver robust anomaly prediction.
- Our method delivers consistently state-of-the-art results across RoadAnomaly [18], Fishyscapes [7] and SMIYC [19], showing strong generalization in diverse datasets.

## II. Related Work

### A. Road Anomaly Segmentation

Road anomaly segmentation aims to localize unexpected obstacles that fall outside the training labels, as evaluated by benchmarks such as Fishyscapes [7] and SMIYC [19].

Uncertainty-based methods segment anomalous regions by deriving anomaly scores from softmax/logit statistics or mask-level confidence. MaxLogits [20], extended from MSP [21], has been applied to driving scenarios through the CAOS [22] benchmark. To alleviate miscalibration in confidence scores, SML [23] normalizes the logit distribution and incorporates boundary refinement for better localization of anomalies. Mask2Anomaly [12] reformulates anomaly segmentation as a mask classification task, achieving improved consistency and recall through mask-level contrastive learning and refinement strategies. Despite these improvements, such methods are prone to false positives, especially in visually distinct but semantically normal regions like trees or the sky.

Outlier exposure-based methods instead leverage auxiliary OOD samples, but their effectiveness depends on the diversity and relevance of the chosen outlier datasets. Maximized Entropy [23] encourages higher-entropy predictions on proxy OOD samples to prevent overconfidence in uncertain regions, while PEBAL [24] employs contrastive learning to differentiate OOD pixels. Similarly, RbA [11] adopts a mask classification paradigm and defines an outlier as any region rejected by all known categories, enabling high-quality OOD segmentation without harming in-distribution performance.

Beyond score-based and outlier-exposure approaches, generative and hybrid methods model inlier feature distributions to obtain likelihood-based OOD scores. DenseHybrid [10] further combines generative modelling of inlier data with discriminative training of negative samples via outlier exposure to enhance OOD segmentation. In addition, image resynthesis-based methods highlight anomalies via reconstruction discrepancies. For example, SynBoost [25] synthesizes images from semantic maps and combines reconstruction cues with uncertainty-based scores.

In this work, we propose VL-Anomaly, which introduces vision-language priors to suppress false positives. To the best of our knowledge, VL-Anomaly is among the first to incorporate multi-modal semantic priors into *road anomaly segmentation*, offering an alternative solution for robust OOD perception in autonomous driving.

### B. Vision-Language Models for Semantic Segmentation

Vision-language models, such as CLIP [13], have demonstrated impressive generalization in open-world tasks by aligning image and text embeddings in a shared semantic space. Recent efforts have integrated VLMs into semantic segmentation to enhance category extensibility and zero-shot capabilities. Methods like OpenSeg [26] and GroupViT [27] leverage VLMs to guide pixel or region-level predictions via text supervision, enabling the segmentation of unseen classes. MaskCLIP [17] further aligns VLM features with segmentation masks to improve visual grounding. However, these works primarily focus on open-vocabulary or zero-shot segmentation within in-distribution domains. More recently, SimCLIP [28] explored the misalignment between high-level language features and fine-grained visual features in industrial anomaly detection, motivating us to exploit VLM-guided semantic priors for road anomaly segmentation to better capture unseen objects. While most existing approaches [17], [26], [27] perform alignment at a single granularity—either pixel-level or mask-level, CoupAlign [29] stands out by coupling the two in a hierarchical manner, which is conceptually closer to our design.

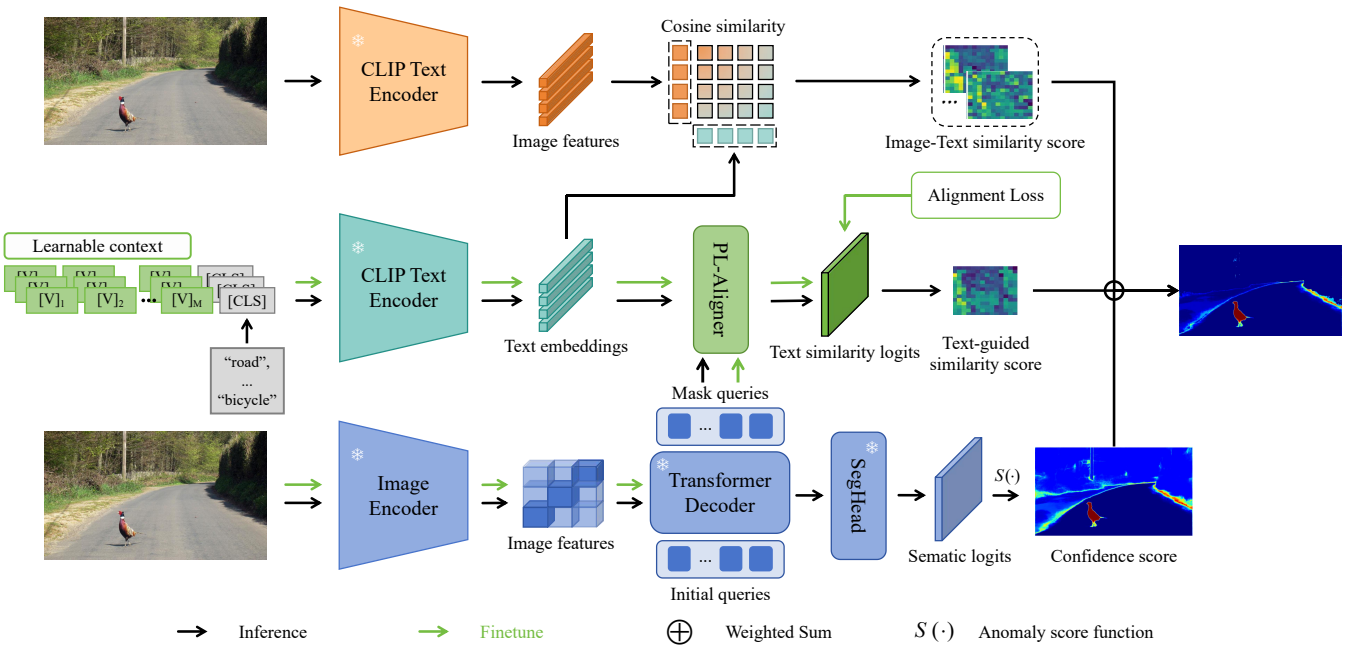


Fig. 2. Overall architecture of VL-Anomaly. The framework integrates a segmentation backbone with CLIP-based vision–language modules. During training, the Prompt Learning-Driven Aligner (PL-Aligner) first performs pixel-level alignment between the backbone's visual features and CLIP text embeddings of known categories, and then further establishes mask-level alignment with the decoder's mask queries. During inference, multi-source scores from the segmentation model outputs, text-guided similarity and CLIP-based image-text similarity are fused to produce robust anomaly segmentation results.

Unlike open-vocabulary segmentation which uses text prompts as classifiers to expand the label space, our goal is not to recognize novel categories but to leverage VLM priors as semantic regularization for suppressing false positives in anomaly segmentation.

## III. Method

In this section, we first outline the problem definition, then review a generic mask-transformer architecture in the context of anomaly segmentation, and finally present our proposed framework and its distinctive components, as demonstrated in fig. 2.

### A. Preliminaries

Formally, let $\mathcal{X} \subset \mathbb{R}^{3\times H\times W}$ denote the space of RGB images, where $H$ and $W$ are the image height and width, $\mathcal{Y} \subset \mathbb{R}^{C_k\times H\times W}$ denote the space of semantic labels assigning each pixel to one of $C_k$ predefined known categories. Given a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{D}$ with $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, the task of anomaly segmentation is to learn a mapping:

$$f : \mathcal{X} \to \mathbb{R}^{H\times W} \tag{1}$$

that produces an anomaly score map for each input image.

In both per-pixel architectures and mask-transformer architectures, $f(\cdot)$ is ultimately used to assign each pixel or predicted mask, which is then compared against a fixed threshold to distinguish OOD regions from in-distribution (ID) ones. While effective in some cases, such confidence frameworks inherently tie the OOD detection performance to the model's learned in-distribution representation, making them susceptible to misclassification in unfamiliar scenes.

To overcome this limitation, we incorporate VLM priors as an external "observer" of semantic knowledge. In contrast to the "insider" perspective of a trained segmentation model, which is confined to task-specific representations, VLMs provide rich open-world semantics that offer an additional viewpoint for distinguishing ID from OOD regions in anomaly segmentation. The following sections detail the technical contributions of our method.

### B. Text Prompt Construction

Unlike typical VLM applications [13] that process a single category description at a time, semantic segmentation requires handling multiple categories simultaneously, as a single image often contains several semantic classes. Compared with manually crafted natural language sentences, adopting the learnable prompt paradigm not only avoids ambiguity and redundancy but also enables automatic adaptation to the segmentation task through joint training, thereby achieving more robust cross-modal alignment [30].

To enable parallel alignment with all known categories, we construct a dedicated prompt for each class $c_i \in \mathcal{C}$, where $\mathcal{C} = \{c_1, c_2, \ldots, c_{C_k}\}$ denotes the set of $C_k$ known semantic classes (e.g., $C_k = 19$ for Cityscapes [31]). This design allows the model to compute class-wise vision-language similarities in a single forward pass, enabling efficient dense prediction. Following the learnable prompt paradigm [30], each prompt adopts a unified context form:

$$\mathbf{p}_i = [\mathbf{V}]_1\,[\mathbf{V}]_2 \ldots [\mathbf{V}]_M\,[\mathrm{CLS}], \tag{2}$$

where $[\mathbf{V}]_i$ $(i \in \{1, \ldots, M\})$ denotes a learnable context token with the same dimension $d$ as the VLM word embeddings,

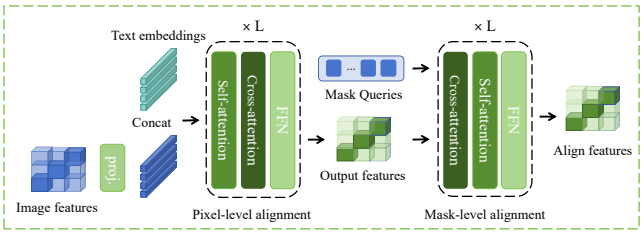


Fig. 3. Architecture of PL-Aligner. The first layer aligns pixel-level visual features from the backbone with text embeddings, while the second layer aligns mask queries from the decoder with the pixel-level features from the first layer to achieve mask-level alignment. Standard operations such as normalization and activation functions are omitted for clarity.

$M$ is the number of context tokens, and [CLS] is the textual name of class $c_i$. The context tokens share the same form across categories and are optimized jointly with the segmentation model to capture task-specific semantics while preserving the VLM's open-world prior.

Given the constructed prompt $\mathbf{p}_i$, the category-specific text embedding is obtained via the VLM's text encoder:

$$\mathbf{t}_i = \text{TextEncoder}(\mathbf{p}_i), \tag{3}$$

where $\mathbf{t}_i \in \mathbb{R}^d$ denotes the text embedding for class $c_i$.

### C. Prompt Learning-Driven Aligner

The feature spaces of the segmentation model and the VLM are inherently misaligned, motivating our learnable prompt-based alignment mechanism. Existing VLM-based segmentation approaches typically adopt a single-granularity strategy, such as pixel-level feature alignment [15] or mask-level alignment based on region features [17]. Inspired by CoupAlign [29], which demonstrates the benefit of combining pixel- and mask-level alignments, we propose a PL-Aligner that jointly enforces fine-grained pixel consistency and structured mask-level semantic alignment, see fig.3.

*a) Pixel-level alignment:* For an input image, the visual encoder first generates a dense feature map, where each spatial location $i$ corresponds to a feature vector $\mathbf{v}_i$. To align these features with the text prompts, the image feature map is re-projected into the same embedding dimension as the text embeddings, concatenated with the prompt features, and processed jointly through attention mechanisms and feed-forward networks. The resulting pixel-aligned features are then aligned with the corresponding text embeddings $\mathbf{t}_{y_i}$ using a pixel-level contrastive loss:

$$\mathcal{L}_{\text{pixel}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}(\mathbf{v}_i, \mathbf{t}_{y_i})/\tau\right)}{\sum_{k=1}^{C_k} \exp\left(\text{sim}(\mathbf{v}_i, \mathbf{t}_k)/\tau\right)}, \tag{4}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, $\tau$ is a learnable temperature parameter, and $y_i$ is the ground-truth label. This stage enforces direct alignment between pixel-level visual features and textual semantics.

*b) Mask-level alignment:* Building upon the pixel-aligned features, we further introduce mask-level alignment after the mask-transformer decoder. Specifically, the pixel-aligned features are used as keys and values, while the decoder mask queries serve as queries. Through attention mechanisms and feed-forward networks, the decoder queries are refined into semantically consistent representations, and then aligned with the text embeddings in the same contrastive manner:

$$\mathcal{L}_{\text{mask}} = -\frac{1}{N_q}\sum_{m=1}^{N_q} \log \frac{\exp\left(\text{sim}(\mathbf{q}_m, \mathbf{t}_{y_m})/\tau\right)}{\sum_{k=1}^{C_k} \exp\left(\text{sim}(\mathbf{q}_m, \mathbf{t}_k)/\tau\right)}, \tag{5}$$

where $\mathbf{q}_m$ denotes the projected mask query and $y_m$ is the ground-truth class label for mask $m$. This stage ensures that the mask-level representations are not only consistent with the pixel-level aligned features but also aligned with the semantic space of the VLM.

*c) Alignment loss:* The final alignment loss combines the segmentation supervision with the two alignment objectives:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{pixel}}\mathcal{L}_{\text{pixel}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}, \tag{6}$$

where $\mathcal{L}_{\text{seg}}$ represents the standard segmentation loss following Mask2Former [12], with $\lambda_{\text{pixel}}$ and $\lambda_{\text{mask}}$ fixed to 0.5 in our finetuning stage.

### D. Multi-source Inference Strategy

Building on the previous section, where the PL-Aligner introduces text awareness during training, we further incorporate external semantic priors from CLIP at inference to better separate ID from OOD regions. To this end, we propose a multi-source inference strategy that integrates three complementary scores: (i) detector confidence, (ii) text-guided similarity, and (iii) CLIP-based image–text similarity. This strategy mitigates the weaknesses of relying on a single source.

*a) Detector confidence:* Inspired by [12], for a mask-transformer architecture, the decoder outputs class scores $\mathbf{c}_m \in \mathbb{R}^{C_k}$ and mask logits $\mathbf{m}_m \in \mathbb{R}^{H\times W}$ for each mask index $m$. The detector confidence map for class $k$ is computed as:

$$S_{\text{conf}} = \max_{k\in\{1,\ldots,C_k\}} \text{softmax}(\mathbf{c}_m)_k \cdot \sigma(\mathbf{m}_m), \tag{7}$$

where $\sigma(\cdot)$ denotes the element-wise sigmoid function.

*b) Text-guided similarity:* We reuse the learned prompt embeddings $\mathbf{t}_k$ obtained during training. The final align features $\mathbf{v}_{align}$ from the PL-Aligner are compared with $\mathbf{t}_k$ to compute the text-guided similarity score:

$$S_{\text{text}}^{(k)} = \text{sim}(\mathbf{v}_{align}, \mathbf{t}_k). \tag{8}$$

This score quantifies the semantic consistency between the aligned visual representation and each learned class-specific prompt embedding.

*c) CLIP-based image-text similarity:* We also compute an image-level similarity score between the input image $x$ and each class prompt $\mathbf{t}_k$ using the frozen CLIP image encoder:

$$S_{\text{img}}^{(k)} = \text{sim}(\text{ImageEncoder}(x), \mathbf{t}_k), \tag{9}$$

which provides a global semantic prior that is independent of the segmentation model's predictions.

TABLE I

EVALUATION ON ROADANOMALY [18], SMIYC-RA21 [19] AND SMIYC-RO21 [19].

| Methods | RoadAnomaly | | | SMIYC-RA21 | | | SMIYC-RO21 | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $FPR_{95}\downarrow$ | AuPRC↑ | AuROC↑ | $FPR_{95}\downarrow$ | AuPRC↑ | AuROC↑ | $FPR_{95}\downarrow$ | AuPRC↑ | AuROC↑ | $FPR_{95}\downarrow$ | AuPRC↑ | AuROC↑ |
| MSP [21] | 71.4 | 15.7 | - | 72.0 | 28.0 | - | 16.6 | 15.7 | - | 53.3 | 19.8 | - |
| Entropy [21] | 68.2 | 15.7 | - | - | - | - | - | - | - | 68.2 | 15.7 | - |
| Mahalanobis [32] | 81.1 | 14.4 | - | 87.0 | 20.0 | - | 13.1 | 20.1 | - | 60.4 | 18.2 | - |
| SML [23] | 70.7 | 17.5 | - | 39.5 | 46.8 | - | 36.8 | 3.4 | - | 49.0 | 22.6 | - |
| JSRNet [33] | **9.2** | **94.4** | - | 43.9 | 33.6 | - | 28.9 | 28.1 | - | 27.3 | 52.0 | - |
| SynBoost [9] | 64.8 | 38.2 | - | 61.9 | 56.4 | - | 3.2 | 71.3 | - | 43.3 | 55.3 | - |
| Max Entropy [34] | 31.8 | 48.9 | - | 15.0 | 85.5 | - | 0.8 | 85.1 | - | 15.9 | 73.2 | - |
| Dense Hybrid [10] | 64.0 | 31.4 | - | 9.8 | 78.0 | - | **0.2** | 87.1 | - | 24.7 | 65.5 | - |
| PEBEL [24] | 44.6 | 45.1 | - | 40.8 | 49.1 | - | 12.7 | 5.0 | - | 32.7 | 33.1 | - |
| ODIN† [35] | 32.9 | 55.0 | 91.4 | 3.6 | 91.5 | 98.5 | 1.1 | 47.8 | 99.2 | 12.5 | 64.8 | 96.4 |
| Mask2Anomaly† [12] | 13.2 | 79.7 | 96.2 | 14.6 | 88.7 | **99.7** | **0.2** | **93.3** | 99.1 | 9.3 | 87.2 | 98.3 |
| VL-Anomaly (Ours) | 12.9 | 79.2 | **96.8** | **3.5** | **95.1** | **99.7** | 0.6 | 91.0 | **99.7** | **5.7** | **88.4** | **98.7** |

*Note:* ↑ indicates higher is better, ↓ indicates lower is better. The best and second best results are **bold** and underlined, respectively.

*d) Score Fusion:* The three scores are combined into a unified anomaly score:

$$S_{\text{final}} = 1 - \max_{k\in\{1,\dots,C_k\}} \left(\alpha \cdot S^{(k)}_{\text{conf}} + \beta \cdot S^{(k)}_{\text{text}} + \gamma \cdot S^{(k)}_{\text{img}}\right), \quad (10)$$

where $\alpha$ ,$\beta$, $\gamma$ are set to 0.7, 0.2 and 0.1, respectively. A higher $S_{\text{final}}$ indicates a greater likelihood of an OOD region.

## IV. EXPERIMENTS

In this section, we first introduce the experiment setup, including the introduction of the anomaly inference datasets, the evaluation metrics and the implementation details. Then, we compare our method with other outstanding baselines. Besides, we provide extensive ablations on both VL-Anomaly modules and the layer design of PL-Aligner.

### A. Setup

**Datasets.** The anomaly inference model is first finetuned on Cityscapes [31] (2,975 training, 500 validation and 1,525 test images), then fine-tuned following [12] using outlier auxiliary datasets generated from MS-COCO [36]. We evaluate on three benchmarks: RoadAnomaly [18], Segment Me If You Can (SMIYC) [19] and Fishyscapes [7]. RoadAnomaly contains 60 Internet-sourced validation images with anomalies of varying scales. SMIYC includes two subsets: RoadAnomaly21 (RA21, 10 validation and 100 test images) and RoadObstacle21 (RO21, 30 validation and 327 test images). Fishyscapes has two subsets: Static (30 validation and 1,000 test images) and Lost & Found (L&F, 100 validation and 275 test images). During inference, only the validation split ground truth is accessible for RoadAnomaly, RA21, RO21 and Fishyscapes; the test ground truth remains unavailable.

**Evaluation metrics.** Following [12], [24], [37], we report the Area Under the Receiver Operating Characteristic curve (AuROC), the False Positive Rate at 95% true positive rate ($FPR_{95}$) and the Area Under the Precision-Recall Curve (AuPRC) for pixel-level evaluation. However, pixel-level metrics may overlook small anomalies and be biased toward large anomalous regions. Therefore, we additionally adopt component-level metrics, including the averaged component-wise F1 score ($F1^*$), the Positive Predictive Value (PPV) and the component-wise Intersection over Union (sIoU).

**Baselines.** In addition to comparing with the results reported in prior works [9], [10], [21], [23], [24], [32]–[34], we further reproduced ODIN† [35] and Mask2Anomaly† [12] for a fairer and more comprehensive comparison. For ODIN [35], the temperature parameter was set to $T = 3.0$, following the implementation provided in the SMIYC benchmark [19] repository. For Mask2Anomaly [12], the refine mask stage was excluded during training. Since the former is a classical work on confidence-based OOD detection and the latter serves as our main baseline for mask-level anomaly segmentation, their inclusion highlights the rationality of our comparison. It is worth noting that all reproduced baselines were applied without any modification to the Mask2Former model architecture, ensuring the fairness of the comparison.

**Implementation details.** Our mask segmentation network is based on Mask2Anomaly [12], which builds upon Mask2Former [5] with a ResNet-50 [38] backbone. During finetuning, the segmentation network and CLIP ViT-L-16 weights are frozen. By default, we employ L = 1 layers of pixel-level and mask-level attention. For prompt learning, we adopt $M = 16$ learnable context tokens per class, each of dimension $d = 512$, initialized from predefined templates via the CLIP's word embeddings, similar to MaskCLIP [17]. The segmentation loss $\mathcal{L}_{\text{seg}}$ follows the standard configuration in Mask2Former, the loss is a weighted sum of classification, Dice [39] and mask losses with coefficients $1.0, 1.0, 20.0$, respectively. The temperature parameter $\tau$ in the contrastive loss is treated as a learnable scalar, initialized to 0.07 following common practice in contrastive learning [13]. For optimization, the model is finetuned using AdamW [40] with an initial learning rate of $1e{-}4$, weight decay 0.05 and batch size 8 for 5,000 iterations. The training images are randomly cropped to $380 \times 760$, and a random scale sampled from from 0.1 to 2.0. During validation, we compute anomaly scores for ID and OOD pixels separately using ground-truth OOD masks, consistent with prior works [12]. All experiments are conducted on a single NVIDIA 3090 GPU.

### B. Evaluation Results

*1) Results on RoadAnomaly and SMIYC-RA21/RO21:* Table I summarizes the results on RoadAnomaly and the

TABLE II

EVALUATION ON FISHYSCAPES STATIC AND FISHYSCAPES LOST & FOUND VALIDATION DATASETS.

| Methods | FS Static | | | FS L&F | | |
|---|---|---|---|---|---|---|
| | $FPR_{95}\downarrow$ | AuPRC↑ | AuROC↑ | $FPR_{95}\downarrow$ | AuPRC↑ | AuROC↑ |
| MSP [21] | 39.8 | 12.9 | - | 44.8 | 1.3 | - |
| Entropy [21] | 39.7 | 15.4 | - | 44.8 | 2.9 | - |
| SML [23] | 20.5 | 52.1 | - | 21.9 | 31.7 | - |
| SynBoost [9] | 15.8 | 43.2 | - | 18.8 | **72.6** | - |
| Max Entropy [34] | 8.6 | 86.5 | - | 35.1 | 29.9 | - |
| Dense Hybrid [10] | 5.9 | 80.2 | - | 3.9 | 47.1 | - |
| PEBEL [24] | **1.7** | **92.4** | - | 7.6 | 44.2 | - |
| ODIN† [35] | 62.4 | 82.3 | 89.4 | 81.1 | 1.4 | 86.3 |
| Mask2Anomaly† [12] | 1.9 | 90.4 | 98.3 | **4.4** | 46.0 | 93.6 |
| VL-Anomaly (Ours) | 2.3 | 90.7 | **98.7** | 8.4 | 69.5 | **96.0** |

*Note:* ↑ indicates higher is better, ↓ indicates lower is better. The best and second best results are **bold** and underlined, respectively.

TABLE III

RESULTS ON THE SMIYC [19] BENCHMARK.

| Methods | SMIYC-RA21 | | | SMIYC-RO21 | | |
|---|---|---|---|---|---|---|
| | sIoU↑ | PPV↑ | $F1^*\uparrow$ | sIoU↑ | PPV↑ | $F1^*\uparrow$ |
| MSP [21] | 15.5 | 15.3 | 5.4 | 19.7 | 15.9 | 6.3 |
| Mahalanobis [32] | 14.8 | 10.2 | 2.7 | 13.5 | 21.8 | 4.7 |
| SML [23] | 26.0 | 24.7 | 12.2 | 5.1 | 13.3 | 3.0 |
| JSRNet [33] | 20.2 | 29.3 | 13.7 | 18.6 | 24.5 | 11.0 |
| Mask2Former [5] | 25.2 | 18.2 | 15.3 | 5.0 | 21.9 | 4.8 |
| Max Entropy [34] | 49.2 | 39.5 | 28.7 | 47.9 | 62.6 | 48.5 |
| SynBoost [9] | 34.7 | 17.8 | 9.9 | 44.3 | 41.7 | 37.5 |
| Dense Hybrid [10] | 54.1 | 24.1 | 31.0 | 45.7 | 50.1 | 50.7 |
| PEBEL [24] | 38.8 | 27.2 | 14.4 | 29.9 | 7.5 | 5.5 |
| Rba [11] | 55.7 | **52.1** | 46.8 | 58.4 | 58.8 | 60.9 |
| Maskomaly [16] | 55.4 | 51.5 | **49.9** | - | - | - |
| Mask2Anomaly† [12] | **60.4** | 45.7 | 48.6 | **61.4** | 70.3 | 69.8 |
| VL-Anomaly(ours) | 59.6 | 50.1 | 48.7 | 61.2 | **73.4** | **70.1** |

*Note:* ↑ indicates higher is better. The best and second best results are **bold** and underlined, respectively.

two subsets of the SMIYC benchmark, RA21 and RO21. VL-Anomaly consistently improves upon existing methods under the same Mask2Anomaly backbone. On RoadAnomaly, it achieves an AuROC of 96.8, which is +0.6 higher than Mask2Anomaly, and reduces $FPR_{95}$ from 13.2 to 12.9. On RA21, VL-Anomaly improves AuPRC by +6.4 compared with Mask2Anomaly. On RO21, it achieves the highest AuROC of 99.7, surpassing Mask2Anomaly by +0.6. When averaged across the three datasets, VL-Anomaly achieves the best overall performance. This aligns with our core idea, as the notable reduction in $FPR_{95}$ and the improvement in AuPRC primarily stem from fewer false positives in semantically normal background regions.

*2) Results on Fishyscapes:* The results on the Fishyscapes benchmark are reported in Table II. On the Static subset, VL-Anomaly achieves an AuROC of 98.7, outperforming Mask2Anomaly by +0.4, while maintaining a competitive $FPR_{95}$ of 2.3. On the more challenging Lost and Found subset, VL-Anomaly shows clear improvements, increasing AuPRC from 46.0 to 69.5 (+23.5) and raising AuROC from 93.6 to 96.0 (+2.4). These results confirm that VL-Anomaly not only excels in simpler settings but also generalizes well to complex scenarios.

TABLE IV

ABLATION ON PIPELINE DESIGN.

| Setting | $FPR_{95}\downarrow$ | AuPRC ↑ | AuROC ↑ | FPS ↑ |
|---|---|---|---|---|
| Baseline | 4.4 | 46.0 | 93.6 | **8.3** |
| + MLP & Hand-crafted prompt | 13.4 | 60.1 | 95.1 | 8.3 |
| → PL-Aligner & $S_{text}$ | 9.2 | 64.0 | 95.1 | 7.9 |
| → Learnable prompt [30] | 9.2 | **69.7** | 95.4 | 7.9 |
| + $S_{img}$ (Full design) | **8.4** | 69.5 | **96.0** | 6.7 |

*Note:* Starting from the baseline, we progressively add prompts, PL-Aligner and multi-source inference components ($S_{text}$:text-guided similarity, $S_{img}$: CLIP-based image-text similarity). ↑ indicates higher is better, ↓ indicates lower is better.

TABLE V

ABLATION ON PL-ALIGNER LAYERS.

| Alignment | Aligner | $FPR_{95}\downarrow$ | AuPRC ↑ | AuROC ↑ |
|---|---|---|---|---|
| Pixel-only | MLP | 13.4 | 60.1 | 93.6 |
| Pixel-only | Cross-attn. [26] | 9.2 | 69.4 | 95.4 |
| Pixel-only | PL-Aligner | 9.1 | 69.4 | 95.6 |
| Mask-only | PL-Aligner | 16.7 | 55.4 | 78.6 |
| Pixel+Mask | PL-Aligner | **8.4** | **69.5** | **96.0** |

*Note:* We compare aligning only with pixel-level, only with mask-level and using both layers (pixel+mask). ↑ indicates higher is better, ↓ indicates lower is better.

*3) Results on SMIYC:* Table III presents the results on the RA21 and RO21 subsets of the SMIYC benchmark. Traditional confidence-based or distance-based approaches such as MSP, Mahalanobis and SML perform poorly, with $F1^*$ scores below 15 on RA21 and below 12 on RO21, showing their limitations in complex driving scenes. Mask-based segmentation frameworks such as Rba and Mask2Anomaly achieve notable improvements, with Mask2Anomaly reaching an sIoU of 60.4 on RA21 and an $F1^*$ of 69.8 on RO21. Building on this strong baseline, VL-Anomaly achieves further gains. On RA21, it achieves the best $F1^*$ of 48.7. On RO21, it improves PPV from 70.3 to 73.4 (+3.1) and $F1^*$ from 69.8 to 70.1 (+0.3), while maintaining a comparable sIoU of 61.2. Overall, VL-Anomaly delivers the best or near-best performance on both RA21 and RO21.

### C. Ablation Study

All the ablation study results reported in this section are from the Fishyscapes lost and found validation dataset [7].

*1) Pipeline design:* We conduct an ablation study in Table IV to evaluate the contributions of each module. Starting from the baseline, adding an MLP-based pixel-level alignment with $S_{\text{text}}$ improves detection performance, although FPR95 increases. Replacing the MLP with our PL-Aligner further enhances results and reduces FPR95. Introducing prompts and moving from hand-crafted to learnable ones brings consistent gains, with AuPRC reaching 69.7 while maintaining 7.9 fps. Finally, incorporating $S_{\text{img}}$ into the multi-source inference delivers the best overall performance at 6.7 fps, which represents only a modest decrease of 1.6 fps compared with the baseline, indicating that our innovations incur minimal efficiency cost.

*2) PL-Aligner layer design:* Table V presents an ablation study on different PL-Aligner configurations. When alignment is performed only at the pixel level, simple MLP projection provides limited improvement, whereas adding cross-

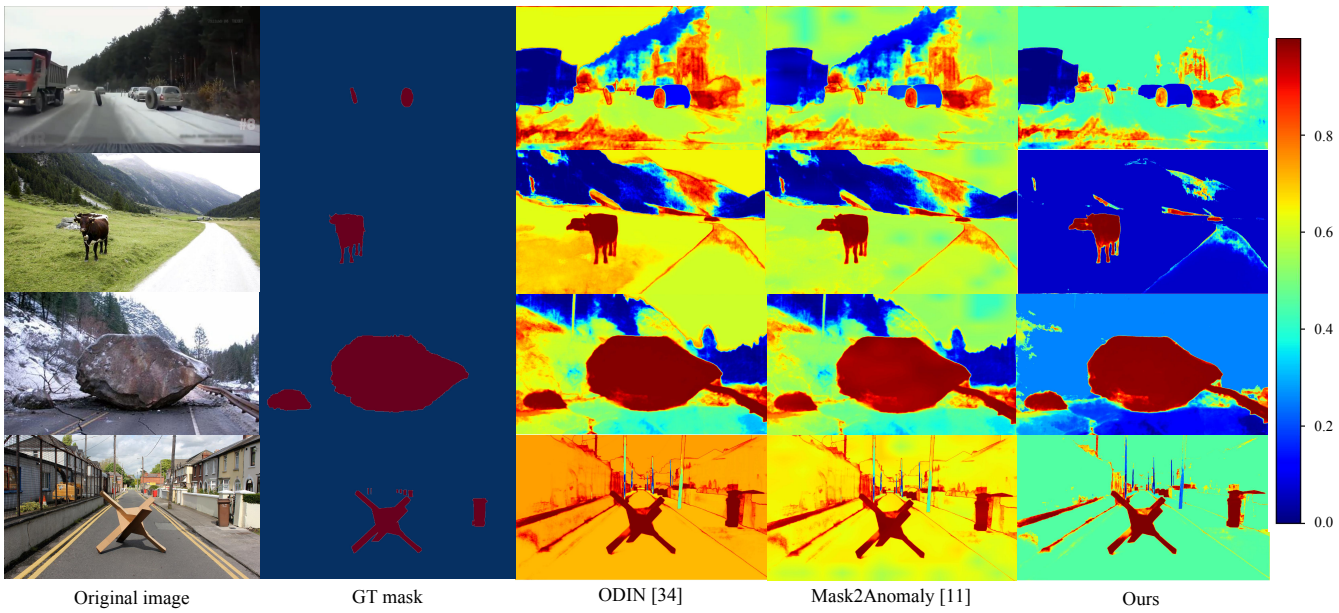


Fig. 4. Qualitative comparison of anomaly segmentation results on the Road Anomaly dataset [18]. We compare the outlier score maps predicted by our method with those generated by MSP [21] and Mask2Anomaly [12], using the same backbone for a fair comparison. For visualization, all scores are normalized to the same range. Our method more effectively suppresses false positives in semantically normal background regions, while competing approaches often yield blurred or spurious activations in these areas.

Fig. 5. Visualization of the similarity between image features and the constructed text prompts. The highlighted areas show where the model associates image regions with specific semantic categories, demonstrating the effectiveness of our prompt learning strategy in guiding cross-modal alignment.

attention markedly reduces FPR95 and boosts AuPRC, confirming the necessity of modeling interactions between visual and textual features. Our pixel-level PL-Aligner achieves comparable gains to cross-attention, but with slightly higher AuROC, showing its effectiveness as a lightweight alternative. By contrast, alignment solely at the mask level performs poorly, indicating that high-level queries alone are insufficient to capture fine-grained semantic cues. The full design, which integrates both pixel-level and mask-level alignment, achieves the best overall results, demonstrating that the two levels provide complementary information: pixel-level alignment anchors fine-grained semantics, while mask-level alignment reinforces structured category consistency.

### *D. Qualitative Results*

Figure 4 shows anomaly maps from ODIN, Mask2Anomaly and VL-Anomaly. Our method effectively suppresses false positives in semantically normal regions such as trees and vegetation, producing cleaner heatmaps with fewer spurious activations. Figure 5 further illustrates the role of our prompt learning strategy. The highlighted areas correspond to normal background categories aligned with text prompts, while anomalous objects emerge as low-intensity regions. This demonstrates that prompt-based cross-modal alignment introduces an additional semantic prior, enabling the model to better disentangle unexpected patterns from the background.

## V. Conclusion

In this work, we presented VL-Anomaly, a vision–language guided anomaly segmentation framework that leverages semantic priors to enhance the segmentation of road anomalies. By introducing the PL-Aligner, we explicitly align visual features with CLIP text embeddings, effectively suppressing false positives in semantically normal regions. Furthermore, our multi-source inference strategy fuses detector confidence, text-guided similarity and CLIP-based Image Similarity to produce more reliable anomaly predictions. Extensive experiments on RoadAnomaly, SMIYC and Fishyscapes benchmarks demonstrate the superiority and generalization ability of our method over state-of-the-art baselines. Although the results are promising, our multi-source inference strategy remains constrained by weights manually adjusted based on multiple datasets, which may limit scalability and automation. Adaptive or data-driven weight learning would offer a more principled alternative. Future work will focus on developing automatic weight optimization strategies to further improve robustness and usability across diverse application settings.

## REFERENCES


[1] J. Chen, J. Lu, X. Zhu, and L. Zhang, "Generative semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 7111–7120.

[2] Z. He, X. Li, J. Tang, S. Qiu, W. Wang, X. Xue, and J. Pu, "Toward camera open-set 3d object detection for autonomous driving scenarios," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 12, pp. 23 190–23 201, 2025.

[3] Y. Xu, J. Cui, F. Cai, Z. Zhu, H. Shang, S. Luan, M. Xu, N. Zhang, Y. Li, J. Cai, *et al.*, "Wam-flow: Parallel coarse-to-fine motion planning via discrete flow matching for autonomous driving," *arXiv preprint arXiv:2512.06112*, 2025.

[4] Y. Liao, S. Kang, J. Li, Y. Liu, Y. Liu, Z. Dong, B. Yang, and X. Chen, "Mobile-seed: Joint semantic segmentation and boundary detection for mobile robots," *IEEE Robotics and Automation Letters*, 2024.

[5] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[6] D. Bo, W. Pichao, and F. Wang, "Afformer: Head-free lightweight semantic segmentation with linear transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[7] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena, "The fishyscapes benchmark: Measuring blind spots in semantic segmentation," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3119–3135, 2021.

[8] D. Bozhinoski, D. Di Ruscio, I. Malavolta, P. Pelliccione, and I. Crnkovic, "Safety for mobile robotic systems: A systematic mapping study from a software engineering perspective," *Journal of Systems and Software*, vol. 151, pp. 150–179, 2019.

[9] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena, "Pixel-wise anomaly detection in complex driving scenes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 918–16 927.

[10] M. Grcić, P. Bevandić, and S. Šegvić, "Densehybrid: Hybrid anomaly detection for dense open-set recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 500–517.

[11] N. Nayal, M. Yavuz, J. F. Henriques, and F. Güney, "Rba: Segmenting unknown regions rejected by all," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 711–722.

[12] S. N. Rai, F. Cermelli, D. Fontanel, C. Masone, and B. Caputo, "Unmasking anomalies in road-scene segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4037–4046.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[14] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[15] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[16] J. Ackermann, C. Sakaridis, and F. Yu, "Maskomaly: Zero-shot mask anomaly segmentation," in *The British Machine Vision Conference (BMVC)*, 2023.

[17] X. Dong, J. Bao, Y. Zheng, T. Zhang, D. Chen, H. Yang, M. Zeng, W. Zhang, L. Yuan, D. Chen, *et al.*, "Maskclip: Masked self-distillation advances contrastive language-image pretraining," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 995–11 005.

[18] K. Lis, K. Nakka, P. Fua, and M. Salzmann, "Detecting the unexpected via image resynthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2152–2161.

[19] R. Chan, K. Lis, S. Uhlemeyer, H. Blum, S. Honari, R. Siegwart, P. Fua, M. Salzmann, and M. Rottmann, "Segmentmeifyoucan: A benchmark for anomaly segmentation," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[20] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," in *International Conference on Machine Learning*. PMLR, 2022, pp. 8759–8773.

[21] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *International Conference on Learning Representations*, 2017.

[22] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," *ICML*, 2022.

[23] S. Jung, J. Lee, D. Gwak, S. Choi, and J. Choo, "Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 425–15 434.

[24] Y. Tian, Y. Liu, G. Pang, F. Liu, Y. Chen, and G. Carneiro, "Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes," in *European Conference on Computer Vision*. Springer, 2022, pp. 246–263.

[25] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena, "Pixel-wise anomaly detection in complex driving scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 16 918–16 927.

[26] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *European conference on computer vision*. Springer, 2022, pp. 540–557.

[27] 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*, 2022, pp. 18 134–18 144.

[28] C. Deng, H. Xu, X. Chen, H. Xu, X. Tu, X. Ding, and Y. Huang, "Simclip: Refining image-text alignment with simple prompts for zero-/few-shot anomaly detection," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1761–1770.

[29] Z. Zhang, Y. Zhu, J. Liu, X. Liang, and W. Ke, "Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 729–14 742, 2022.

[30] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision (IJCV)*, 2022.

[31] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[32] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, 2018.

[33] T. Vojir, T. Šipka, R. Aljundi, N. Chumerin, D. O. Reino, and J. Matas, "Road anomaly detection by partial image reconstruction with segmentation coupling," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 651–15 660.

[34] R. Chan, M. Rottmann, and H. Gottschalk, "Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation," in *Proceedings of the ieee/cvf international conference on computer vision*, 2021, pp. 5128–5137.

[35] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *International Conference on Learning Representations*, 2018.

[36] 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*. Springer, 2014, pp. 740–755.

[37] Y. Liu, C. Ding, Y. Tian, G. Pang, V. Belagiannis, I. Reid, and G. Carneiro, "Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1151–1161.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016.

[40] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.