Dual-Stream Adapters for Open-Set Segmentation in Driving Scenes

Dual-Stream Adapters for Open-Set Segmentation in Driving Scenes

Shyam Nandan Rai[1]
shyam.rai@polito.it
Massimiliano Mancini[2]
massimiliano.mancini@unitn.it
Barbara Caputo[1]
barbara.caputo@polito.it
Carlo Masone[1]
carlo.masone@polito.it

[1] Politecnico di Torino, Italy
[2] University of Trento, Italy

Abstract

The task of segmenting novel categories in road scenes, often referred to as anomaly segmentation, has been recently addressed with great success by using mask-based architectures, but their efficacy is dependent on fine-tuning large transformer backbones. In this work, we design a specialized adapter for this task, which makes it possible to leverage even large backbones without re-training them. The key feature of our adapter is the separation of the adapted features in two streams, one specialized on the known categories (in-distribution) and the other that captures the characteristics of out-of-distribution categories. The out-of-distribution features adaptation is supervised by using synthetic negative data generated by a normalizing flow process. This dual-stream architecture allows to better disentangle features for known and unknown categories, preserving in-distribution performance while enabling direct and more accurate anomaly segmentation with fewer false positives. Experiments show that dual-stream adapters outperform previous methods while reducing training parameters by 38%.

1 Introduction

Semantic segmentation plays a critical role in autonomous and assisted driving, enabling the fine-grained scene understanding necessary for advanced functionalities like path planning and obstacle avoidance. Models used in this context are typically trained with a predefined, fixed set of relevant categories (e.g., pedestrian etc.), a *closed-set* assumption that rarely holds in real-world driving scenarios. Therefore, it is crucial that these models recognize instances of unknown categories [1, 3, 15, 19]. Failure to do so results in erroneous classifications, potentially leading to catastrophic misinterpretations and hazardous decisions.

In this context, a promising line of research has shown that it is advantageous to resort to mask-based architectures to identify new categories (often referred to as anomalies [4]), by reasoning on whole masks rather than individual pixels [1, 37, 39, 40]. However, the

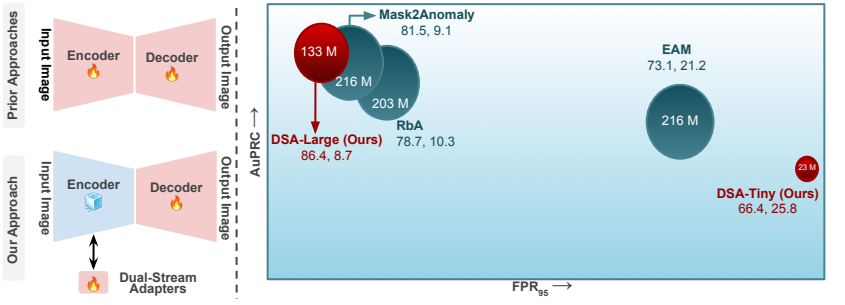

Figure 1: **Left:** State-of-the-art methods [20, 37, 40] rely on large backbones like Swin-L [35], requiring huge trainable parameters. We propose **Dual-Stream Adapter** with a frozen pre-trained backbone for efficient anomaly segmentation. **Right:** DSA-Large surpasses top methods using 38% fewer trainable parameters. Tuples show average AuPRC and FPR_{95}; circles indicate trainable parameters (in millions).

improvements achieved by these architectures are largely dependent on the training of large visual transformer backbones [13, 35] with a significant number of training parameters and consequently a substantial training cost (cf. Fig. 1). Another effective strategy is to combine an anomaly scoring function computed on the closed-set segmentation model (*e.g.* using MSP [24] or RbA [37]) with an additional explicit prediction of the anomalies, *e.g.* obtained by extending the classification head with a negative object class [11] or by adding a separate dedicated decoder [43].

In this work, we revisit these ideas –mask-based anomaly segmentation and the combination of implicit and explicit anomaly predictions– with the goal of developing a method for open-set segmentation of driving scenes that is both more efficient (in terms of training parameters) and effective than the current state-of-the-art. Our solution is inspired by vision transformer (ViT) adapters, which have shown to enable the use of large pretrained models for dense prediction tasks without requiring a full fine-tuning [6]. However, differently from the vanilla ViT adapter, here we seek to design an adapter that not only adjusts the learned features to make dense predictions about the in-distribution (ID) categories, but also disentangles ID and out-of-distribution (OOD) concepts. Our solution to this problem, named ***Dual-Stream Adapter***, learns to extract features for ID and OOD categories, making it possible to obtain more nuanced and straightforward-to-predict anomalies (implicitly and explicitly) than using a single set of shared features. To achieve this, the Dual-Stream Adapter uses a symmetric structure for ID and OOD features, with two key components:

- an *anomaly prior module* that captures local context using a convolutional stem and provides an initial set of in-distribution and out-of-distribution features;
- *dual-stream feature refinement modules* that iteratively refine these initial features and separately combine them with the general representations of a frozen ViT backbone.

The two streams of features are trained using supervision of both ID and OOD concepts, with a contrastive loss to separate and disentangle them. Although this joint training requires data containing negative concepts, we forego using real negative data and rely just on synthetic generated negative data and, when present, void/background masks in the inlier training data (*e.g.* in Cityscapes). We perform extensive experiments on anomaly segmentation benchmarks (Fishyscapes [3], Segment Me If You Can [4], Road Anomaly [34]), show-

ing Dual-Stream Adapter achieves the best results among all anomaly segmentation methods with 38% lower training parameters, when compared to the best baseline method (cf. Fig. 1).

2 Related Work

Road Obstacle Segmentation is synonymous with open-set segmentation, aiming to detect categories unseen during training. Early solutions adapted methods from unsupervised anomaly segmentation, such as maximum softmax probability prediction [24], deep ensembles [14], Bayesian deep learning [15, 36, 45], class-logits reasoning [25, 27], and image re-synthesis discrepancy [34]. However, recent methods specifically developed for open-set segmentation in driving scenes mostly provide explicit supervision to the model by training on mixed images that contain instances of novel categories (*negative crops*) pasted on them [2, 10, 11, 12, 30]. These negative crops are generally sampled from the MS-COCO [33] or ADE-20K [51] datasets, which contain objects absent in in-distribution datasets. Alternatively, when appropriate negative data may be unavailable, synthetic negatives may be used effectively, as demonstrated by [11, 17, 18] which use a normalizing flow as a generative process. Many state-of-the-art road obstacle segmentation methods [1, 20, 37, 39, 40, 49] use mask-based architectures [7, 8] that treat segmentation as set prediction via binary masks with class labels. Methods like RBA [37], EAM [20], Mask2Anomaly [39, 40], and Maskomaly [1] improve anomaly segmentation by aggregating scores over predicted masks instead of individual pixels. However, they rely on large Transformer backbones (*e.g.* ViT [13], Swin [35]) with high training parameters. Recent works [11, 43] also introduce explicit anomaly mask prediction by extending closed-set models. For example, UNO [11] adds a new class, while ContMAV [43] employs a second decoder. In comparison to prior works, we propose an adapter tailored for open-set segmentation in driving scenes, preserving the benefits of mask-based architectures without retraining the backbone. Moreover, it generates separate ID and OOD feature streams, maintaining strong performance on known classes while explicitly predicting anomaly masks. In contrast, [11, 43] has a shared classifier for both known and unknown classes segmentation.

Adapters. Learning universal representations that can be specialized to multiple tasks and domains is quintessential in deep learning. Adapters [41, 42] have emerged as an effective and parameter-efficient solution to this problem. In computer vision, the emergence of large pre-trained models [28, 38] is making the utilization of adapters quite compelling. In fact, fine-tuning these models for possibly many downstream tasks is inefficient not only in terms of required training resources, but also in terms of deployed resources, as a device that must perform multiple tasks would require multiple specialized copies of the same model. Moreover, it was also demonstrated that fine-tuning these large vision models could distort the pre-trained features [29] and lead to poor generalization. This has inspired a few works to use a vanilla ViT architecture [13] and adapt it to various vision tasks [31, 32]. Most notable in this sense is the ViT-Adapter [6], which integrates vision-specific inductive biases into the plain ViT, reaching comparable performance to recent vision-specific transformer variants on several tasks but training fewer parameters. Unlike the general-purpose ViT-Adapter [6], our Dual-Stream Adapter is designed for open-set segmentation by explicitly separating in-distribution and anomalous features.

3 Preliminaries

3.1 Problem Setting

We consider a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{D}$, where $x_i \in \mathcal{X} \subset \mathbb{R}^{3\times H\times W}$ is an RGB image, and $y_i \in \mathcal{Y} \subset \mathbb{N}^{|\mathcal{Z}|\times H\times W}$ is the corresponding pixel-wise semantic label mask over a set of pre-defined categories $\mathcal{Z}$. Here, H and W denote the height and width, respectively, of each sample of $\mathcal{X}$ and $\mathcal{Y}$. The goal is to train a model that outputs both a semantic segmentation mask and a binary anomaly mask indicating pixels that do not belong to any known class. Traditionally, anomaly detection is treated as a per-pixel classification task [19, 44], involving a function $f_\theta : \mathcal{X} \to \mathbb{R}^{|\mathcal{Z}|\times H\times W}$ and then applying a non-parametric scoring function $s : \mathbb{R}^{|\mathcal{Z}|\times H\times W} \to [0,1]^{H\times W}$. Recent methods [1, 20, 37, 39, 40, 49] instead use the Mask2Former architecture [8], moving from per-pixel to mask-based classification. We next summarize this reformulation, as it forms the foundation of our solution.

3.2 Mask2Former-based anomaly segmentation

Mask2Former [8] is a universal architecture that approaches segmentation as a direct set prediction problem. It groups pixels into N segments by predicting N binary masks and their corresponding category labels (where N is a hyper-parameter). The architecture consists of three main components: (i) an *encoder* that extracts the features from the input image, (ii) a *pixel-decoder* that generates high-resolution per-pixel embeddings from the encoder's low-resolution features, and (iii) a *transformer-decoder* that operates on image features to process object queries. For brevity, we do not delve into a detailed description of these components. Interested readers can refer to [8] for a comprehensive explanation of the Mask2Former architecture. Additionally, for the sake of compactness and readability, we refer pixel decoder and transformer decoder collectively as the *decoder* (see Fig. 2, Fig. 1-left). Given this architecture, the aim is to learn parameters θ of a function f_θ, composed of an encoding function h_{θ_e} and a decoding function g_{θ_d} with $\theta = [\theta_e, \theta_d]$, that maps an image to a set of binary masks $M \in \mathbb{R}^{N\times(H\times W)}$, and category labels $C \in \mathbb{R}^{N\times|\mathcal{Z}|}$ formulated as:

$$f_\theta : \mathcal{X} \to (\mathbb{R}^{N\times(H\times W)}, \mathbb{R}^{N\times|\mathcal{Z}|}), \quad x \mapsto g_{\theta_d} \circ h_{\theta_e}(x) = (M, C) \tag{1}$$

While the masks M and class scores C associate pixels to the known classes from $\mathcal{Z}$, it was shown in [40] that a binary anomaly mask can be predicted implicitly from M and C through a non-parametric scoring function $s_{impl} : (\mathbb{R}^{N\times(H\times W)}, \mathbb{R}^{N\times|\mathcal{Z}|}) \to \mathbb{R}^{H\times W}$. For example, using MSP [24], the scoring function becomes

$$s_{impl}(M, C) = 1 - \max_{\mathcal{Z}} \left(\texttt{softmax}(C)^T \cdot \texttt{sigmoid}(M)\right) \tag{2}$$

Without loss of generality, other scoring functions can also be used, *e.g.*, [20, 37] and have been crucial in achieving state-of-the-art results, particularly with large pre-trained encoders [1, 20, 37, 39, 40, 49]. However, fine-tuning such encoders is computationally expensive and may distort pre-trained features [29], reducing effectiveness for other downstream tasks. To alleviate this issues while maintaining state-of-the-art performance, we propose freezing the pre-trained encoder and inserting an adapter module [26, 42] with learnable parameters θ_a, such that $\|\theta_a\| << \|\theta_e\|$ (see Fig. 1-left). Namely, we now seek a map

$$f_\theta : \mathcal{X} \to (\mathbb{R}^{N\times(H\times W)}, \mathbb{R}^{N\times|\mathcal{Z}|}), \quad x \mapsto g_{\theta_d} \circ h_{\tilde{\theta}_e, \theta_a}(x) = (M, C) \tag{3}$$

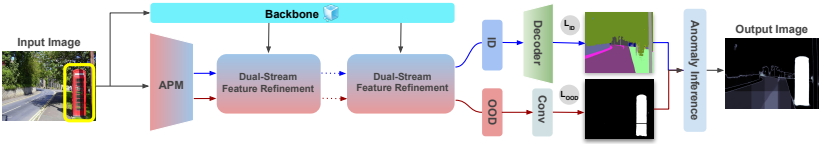

Figure 2: **Dual-Stream Adapter:** Our model includes a frozen ViT backbone, a decoder, and a dual-stream adapter with two components: the anomaly prior module (APM) (Sec. 4.1) and the dual-stream feature refinement (Sec. 4.2), extracting in-distribution and out-of-distribution features. The decoder (Mask2Former decoder) and Conv (single convolutional layer) are fused at inference to produce the anomaly map (Sec. 4.2). The yellow box indicates an anomaly. $\mathcal{L}_{ID}$ and $\mathcal{L}_{OOD}$ are used to train the ID and OOD streams (Sec. 4.3).

where $\bar{\theta}_e$ denotes the frozen parameters. We model this adapter after the ViT-Adapter [6], as explained in the next section.

4 Dual Stream Adapter

The Dual-Stream Adapter (DSA) architecture (see Fig. 2) builds on ViT-Adapter [6]. It incorporates an anomaly prior module for initial ID and OOD feature extraction and a chain of dual-stream refinement that iteratively enhances these features using the frozen ViT. In the following subsections, we discuss these technical novelties in detail. For clarity, adapter dependencies on θ_a are omitted.

4.1 Anomaly Prior Module

The first part of the DSA is the anomaly prior module (see Figs. 2 and 3). Our module shares a similar foundation with the ViT-Adapter's Spatial Prior Module, leveraging a ResNet convolutional stem [23] to extract D-dimensional feature maps at different spatial resolutions: ($\mathcal{F}_8 \in \mathbb{R}^{D\times\frac{H}{8}\times\frac{W}{8}}$, $\mathcal{F}_{16} \in \mathbb{R}^{D\times\frac{H}{16}\times\frac{W}{16}}$ and $\mathcal{F}_{32} \in \mathbb{R}^{D\times\frac{H}{32}\times\frac{W}{32}}$). The intuition of this design is that the convolutional stem complements the vision transformer by capturing local context and providing translation equivariance [47]. Each of the three feature tensors is further augmented with an additive and learnable *level encoding*. Formally, for a generic feature scale $\mathcal{F}_i$, it is given as $\bar{\mathcal{F}}_i = \mathcal{F}_i + \delta_i$. Where, $\delta_i \in \mathbb{R}^{D\times\frac{H}{i}\times\frac{W}{i}}$ is the corresponding level encoding. All these features are finally flattened and concatenated together. We diverge from ViT-Adapter by introducing an explicit specialization mechanism. Features are processed in two separate streams with distinct level encodings, enabling them to specialize on either ID or OOD object characteristics (see Fig. 3). Formally, for a generic feature scale $\mathcal{F}_i$, we have

$$\bar{\mathcal{F}}_{i,\text{id}} = \mathcal{F}_i + \delta_{i,\text{id}}, \quad \bar{\mathcal{F}}_{i,\text{ood}} = \mathcal{F}_i + \delta_{i,\text{ood}} \tag{4}$$

where $\delta_{i,\text{id}}, \delta_{i,\text{ood}} \in \mathbb{R}^{D\times\frac{H}{i}\times\frac{W}{i}}$ are the corresponding level encodings, each of them obtained as a distinct D-dimensional learnable vector broadcasted along the spatial coordinates. The level encoding here acts as a channel-wise bias individually learned for in-distribution and out-of-distribution features, which allows to retain information about their resolution level within the feature tokens. The feature tensors of the two streams are finally flattened and concatenated separately, yielding two arrays of D-dimensional feature tokens, i.e.,

$$\bar{\mathcal{F}}_{8,\text{id}}, \bar{\mathcal{F}}_{16,\text{id}}, \bar{\mathcal{F}}_{32,\text{id}} \mapsto \mathcal{F}^1_{\text{id}} \in \mathbb{R}^{(\frac{HW}{8^2}+\frac{HW}{16^2}+\frac{HW}{32^2})\times D}, \quad \bar{\mathcal{F}}_{8,\text{ood}}, \bar{\mathcal{F}}_{16,\text{ood}}, \bar{\mathcal{F}}_{32,\text{ood}} \mapsto \mathcal{F}^1_{\text{ood}} \in \mathbb{R}^{(\frac{HW}{8^2}+\frac{HW}{16^2}+\frac{HW}{32^2})\times D} \tag{5}$$

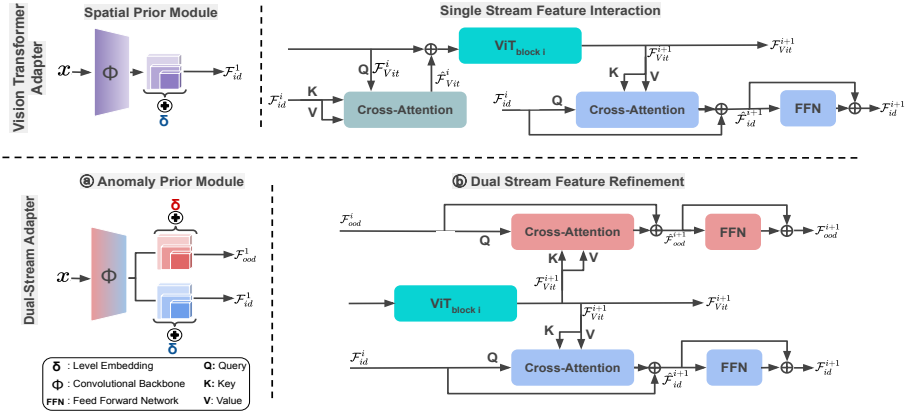

Figure 3: **Above:** Shows the spatial prior module and adapter architecture present in the existing vision adapters architecture [6]. **Below:** We present novel components of our dual-stream adapter. ***(a)Anomaly Prior Module (Sec. 4.1):*** learn to extract the initial ID and OOD features from the input image. ***(b)Dual-Stream Feature Refinement (Sec. 4.2):*** takes in the initial ID and OOD features from the anomaly prior module. The features are refined by passing through a set of anomaly adapters and augmenting with the ViT features.

With the anomaly prior module, the architecture can learn strong prior knowledge about the ID and OOD features, as demonstrated by Tab. 4 (a).

4.2 Dual Stream Feature Refinement

The original ViT-Adapter [6] uses an interaction module consisting of cross-attentions to combine the spatial information from the prior module and the features from the frozen ViT. For our problem, we want to combine the frozen ViT features with the spatial in-distribution (ID) and out-of-distribution (OOD) information, while at the same time disentangling them. Therefore, we design a module based on cross-attention with a symmetric structure to handle the dual streams information. Each *dual-stream feature refinement* module consists of two symmetric streams: one for the ID features and the other for the OOD features. Without loss of generality, let us consider the ID stream for the i-th dual-stream feature refinement module. To combine the spatial ID prior with the ViT features, we use a cross-attention layer, where the keys and values are given from features produced by the $(i+1)$-th ViT block, i.e., $\mathcal{F}^{i+1}_{\text{Vit}}$. The spatial features $\mathcal{F}^{i}_{\text{id}}$ are used as queries to imbue relevant multi-scale features from $\mathcal{F}^{i+1}_{\text{Vit}}$ into the tokens of $\mathcal{F}^{i+1}_{\text{id}}$. The same sequence of operations is implemented for the OOD stream. The whole process is summarized as:

$$\textit{i.d. extractor}\quad \mathcal{F}^{i+1}_{\text{id}} = \hat{\mathcal{F}}^{i+1}_{\text{id}} + \texttt{FFN}(\hat{\mathcal{F}}^{i+1}_{\text{id}}), \quad \hat{\mathcal{F}}^{i+1}_{\text{id}} = \mathcal{F}^{i}_{\text{id}} + \texttt{Attention}(\underbrace{\mathcal{F}^{i}_{\text{id}}}_{Q}, \underbrace{\mathcal{F}^{i+1}_{\text{ViT}}}_{K}, \underbrace{\mathcal{F}^{i+1}_{\text{ViT}}}_{V}) \tag{6}$$

$$\textit{o.o.d. extractor}\quad \mathcal{F}^{i+1}_{\text{ood}} = \hat{\mathcal{F}}^{i+1}_{\text{ood}} + \texttt{FFN}(\hat{\mathcal{F}}^{i+1}_{\text{ood}}), \quad \hat{\mathcal{F}}^{i+1}_{\text{ood}} = \mathcal{F}^{i}_{\text{ood}} + \texttt{Attention}(\underbrace{\mathcal{F}^{i}_{\text{ood}}}_{Q}, \underbrace{\mathcal{F}^{i+1}_{\text{ViT}}}_{K}, \underbrace{\mathcal{F}^{i+1}_{\text{ViT}}}_{V}) \tag{7}$$

where `Attention` denotes the cross-attention block and `FFN` is a feed-forward network as in the original ViT-Adapter [6]. Besides its dual structure, our module differs from

ViT-Adapter's interaction module for its lack of an injector cross-attention (see Fig. 3) module. This design choice is empirically motivated, as we found that the anomaly segmentation results are similar with and without the injector (cf. supplementary). However, removing the injector cross-attention results in a more streamlined module, with fewer training parameters. In summary, this structure helps to refine and specialize the features obtained from anomaly prior to improve anomaly segmentation while retaining a good performance on the in-distribution samples.

4.3 Training and inference

Decoding ID and OOD features. ID features from the Dual Stream Adapter are decoded by Mask2Former into binary masks M and class scores C (eq. (3)), producing the final segmentation mask over known classes $\mathcal{Z}$ (Fig. 2). OOD features pass through a point-wise convolution, followed by sigmoid normalization and resizing to the input dimensions, yielding the **explicit anomaly score map** s_{expl} (Fig. 2). This explicit map is obtained by supervising the model with synthetic negative data, as detailed next.

Streams supervision and loss function. The adapter generates two feature streams: one for ID and another for OOD categories. While ID features can be directly supervised, OOD supervision is challenging due to the lack of negative data. A simple solution to this problem is to use void/background pixels as OOD samples. However, since void/background pixels are not always available in the training data, a more widely applicable alternative is to use synthetic negative patches pasted onto the training data. This technique has been widely used in this setting [11, 17, 18, 21]. Following [17, 21], we use a pre-trained normalizing flow model N_f (pre-trained on Pascal-COCO) to generate synthetic negatives. Given Gaussian noise z_g, N_f produces a patch x_z, which is pasted onto a training image x to obtain the augmented image x^a. Formally, this can be expressed as:

$$x^a = (1 - b_m) \cdot x + b_m \cdot \texttt{pad}(x_z), \quad x_z = N_f(z_g) \tag{8}$$

where b_m is a binary mask with 1 at x_z's pixel locations, and pad applies zero-padding to match x's size. During training, x^a is labeled differently per stream: x_z pixels are treated as anomalies for the OOD stream (convolutional output in Fig. 2) and as background for the ID stream (decoder output in Fig. 2). We apply normalizing flow patches with probability p_{nf} and treat void/background pixels as anomalies with probability p_{bg}. Using data from eq. (8), we train the ID stream with Mask2Former's losses: $\mathcal{L}_{ID} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{dice}\mathcal{L}_{dice}$, and the OOD stream with the mask contrastive loss $\mathcal{L}_{OOD}$ [40]. The total loss is $\mathcal{L}_{DSA} = \mathcal{L}_{ID} + \mathcal{L}_{OOD}$.

Inference with ID and OOD features. At inference, we use the implicit score s_{impl} from eq. (2) and the explicit anomaly map s_{expl} from the OOD stream (Fig. 2). The final anomaly map is computed as a weighted sum: $s = s_{impl} + \lambda_s \cdot s_{expl}$, with $\lambda_s = 0.01$ for DSA-Large and 0.1 for DSA-Tiny.

5 Experiments

Implementation details:. We evaluate two configurations of our model, using Dual Stream Adapters: Tiny (DSA-Tiny) and Large (DSA-Large), which utilize ViT-Tiny and a ViT-Large backbone, respectively. We evaluate all models on Road Anomaly [34], Fishyscapes (FS) [3]

Methods	Road Anomaly		SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Average	
	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓
Fine-Tuned	28.4 | 39.1	85.4 | 76.4	57.4 | 67.2	82.2 | 50.8	63.6 | 10.7	2.6 | 29.8	44.1 | 66.2	33.1 | 28.3	29.5 | 89.1	88.6 | 12.3	44.6 | 54.4	58.3 | 39.5
Side Adapters [28]	23.1 | 44.8	85.2 | 62.1	60.3 | 61.1	77.3 | 45.9	29.3 | 53.8	12.1 | 41.6	45.1 | 47.2	**23.6** | 40.9	25.2 | **88.9**	88.3 | 15.9	36.6 | 59.1	57.3 | 41.2
ViT Adapters [6]	26.0 | 37.5	88.3 | 78.5	57.1 | 60.7	90.7 | 24.9	60.9 | 79.5	7.5 | 16.0	42.5 | 44.9	27.0 | 42.6	**31.5** | 80.6	33.1 | 27.3	43.4 | 60.6	49.3 | 37.3
DSA-Tiny (Ours)	**30.9** | **48.9**	**83.1** | **57.1**	**60.3** | **68.8**	**24.2** | **21.1**	**89.9** | **85.4**	**1.2** | **3.4**	**50.3** | **54.3**	65.7 | **37.7**	30.9 | 87.5	**25.0** | **10.3**	**52.7** | **68.9**	**39.8** | **25.9**

Table 1: **Quantitative comparison of adapters:** We present the performance comparison between baseline vision adapters (Side Adapters [28], ViT Adapters [6]) and our proposed DSA-Tiny adapters. Each paired entry in the table represents performance: without outlier supervision | with outlier supervision. The best results are in **bold**.

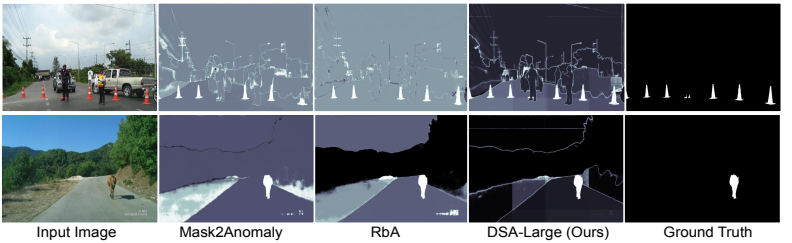

Figure 4: **Qualitative Results**: DSA-Large provide better and crisper anomaly masks w.r.t. other baselined methods. Anomalies are represented in white. The samples are shown for SMIYC RA-21 and SMIYC RO-21. Please refer to the supplementary for more results.

and Segment Me If You Can (SMIYC) benchmark [4]. We employ AuPRC and FPR_{95} as evaluation metrics. Please refer to the supplementary for further details.

Baselines and fairness:. We compare our method against a wide array of baselines, including both mask-based and per-pixel approaches. For all mask-transformer methods [1, 20, 37, 40], we use Swin-L [35]. Further experimental details are in the supplementary. Table 2 lists all methods. To ensure fairness, all models, including ours, are trained on only Cityscapes [9]. As outlier exposure step, we adopt the same fine-tuning protocol as [40]. Additional information on outlier supervision is given in the supplementary.

5.1 Main Results

Comparison to other adapters.
Setting: We evaluate the dual-stream adapter with other ViT adapters by using the same ViT-Tiny backbone and mask decoder. During outlier supervision, all models are trained with mask-contrastive loss [40]. Aside from the adapters themselves, all other implementation details are the same used by DSA-Tiny.
Discussion: Table 1 presents the performance of existing vision adapters compared to our DSA-Tiny, with and without outlier exposure. On average, DSA-Tiny shows significant performance improvement over existing vision transformer adapters by achieving higher AuPRC values and lower FPR_{95} scores, highlighting the effectiveness of the anomaly prior and the dual stream feature refinement block. The qualitative comparison shown in supplementary material supports these numerical results, showing that the dual stream adapters not only is more capable to segment the anomaly, but it also produces fewer false positives. Notably, fine-tuning the entire architecture yields lower performance compared to our approach. This is in line with the findings of [29] that fine-tuning leads to distorting the pre-trained features, consequently reducing anomaly detection performance.

Methods	Real Negatives	SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly		Average	
		AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓
Max Softmax [23]	✗	27.9	72.0	15.7	16.6	4.5	40.5	19.0	23.9	15.7	71.3	16.6	44.9
Mahanalobis [31]	✗	20.0	86.9	20.9	13.0	56.3	11.2	27.3	11.7	14.3	81.0	27.8	40.8
Image Resynthesis [20]	✗	52.2	25.9	37.7	4.7	5.7	47.7	8.0	62.7	-	-	25.9	35.3
MC Dropout [20]	✗	28.8	69.4	4.8	50.3	12.2	32.8	42.1	13.2	-	-	22.0	41.4
Learning Embedding [8]	✗	37.5	70.7	0.8	46.3	4.1	22.3	43.5	16.8	-	-	53.1	42.6
SML [20]	✗	46.8	39.5	3.4	36.8	36.5	14.5	48.6	16.7	17.5	70.7	22.0	41.4
ATTA [20]	✗	67.0	31.5	76.4	2.8	**65.5**	**4.4**	**93.6**	**1.1**	59.0	33.4	72.3	14.6
FlowEneDet [20]	✗	36.7	77.8	73.7	0.9	56.1	**3.8**	66.6	8.9	-	-	58.3	22.9
Maximized Entropy [8]	✗	85.4	15.0	85.0	**0.7**	40.8	37.2	72.4	12.9	48.8	31.7	66.5	19.5
SynBoost [20]	✓	56.4	61.8	81.7	4.6	40.9	34.4	48.4	47.7	38.2	64.7	21.5	39.0
JSRNet [20]	✓	33.6	43.8	28.0	18.5	0.2	69.3	1.4	60.4	**94.4**	**9.2**	31.5	40.2
Dense Hybrid [20]	✓	77.9	9.8	87.0	**0.2**	63.8	6.1	60.0	4.9	31.3	63.9	64.0	17.0
PEBEL [20]	✓	49.1	40.8	4.9	12.6	59.8	6.4	82.7	6.8	45.1	44.5	48.3	22.2
EAM [20]	✓	76.3	93.9	66.9	17.9	52.0	20.5	87.3	**2.1**	29.8	54.9	62.5	37.9
Maskomaly [8]	✓	**93.4**	**6.9**	-	-	-	-	69.5	14.4	16.3	73.1	59.7	31.5
Mask2Anomaly [20]	✓	90.5	9.8	72.6	8.0	**66.8**	**4.6**	**91.2**	6.0	75.4	**9.8**	**79.3**	**7.7**
RBA [20]	✓	**92.3**	**7.2**	**92.4**	**0.3**	50.4	10.8	69.5	6.0	**83.5**	18.7	**77.6**	**8.6**
UNO [20]	✓	91.3	8.8	**89.4**	1.3	35.0	42.8	82.3	12.2	54.1	17.0	70.5	16.4
DSA-Large (Ours)	✓	**92.0**	**4.3**	**85.5**	5.3	**73.9**	11.4	**89.9**	**2.0**	**80.1**	**11.4**	**84.2**	**6.8**

Table 2: **Quantitative Results:** We observe that on average our DSA-Large model obtains the best results among the baseline anomaly segmentation methods. The first, second and third best results are reported in **green**, **orange** and **red**, respectively.

	Mask2Former [8]	ViT-Adapter [8]	RbA [20]	EAM [20]	DSA-Large (Ours)
without Outlier supervision	83.37	83.30	82.85	83.27	**84.93**
with Outlier supervision	-	-	82.25	82.16	**84.62**

Table 3: **In-distribution performance**: Results on the Cityscapes validation set, for the vanilla Mask2Former and recent anomaly segmentation SOTA based on it. DSA-Large retains the best in-distribution performance among these models. Best results are in **bold**.

Comparison to other anomaly segmentation methods.

Anomaly Segmentation Performance: Table 2 presents the performance of per-pixel and masked-based anomaly segmentation methods. We observe that in average DSA-Large obtains the best results among the recent SOTA mask-based architectures such as Mask2Anomaly [40] and RbA [37], although it appears that the results of the various methods vary depending on the datasets. The visual comparison in Fig. 4 shows that DSA-Large is more effective at segmenting the anomalies, with minimal false positives.

Semantic Segmentation Performance: We show segmentation performance on in-distribution data, as improving OOD detection can sometimes degrade ID performance, posing as a trade-off requiring separate models. Table 3 shows results on the Cityscapes validation set for Mask2Former and related anomaly segmentation methods. DSA-Large achieves the best ID accuracy, with or without outlier supervision, highlighting the effectiveness of our dual-stream adapter. Qualitative results are in the supplementary material.

5.2 Ablation Study

We conducted all ablation experiments with the DSA-Tiny architecture on the SMIYC RO-21 dataset, unless specified. Additional ablation are in the supplementary material.

Component-wise ablation. Table 4(a) shows ablations of dual-stream adapter on SMIYC RO-21. We find that removing anomaly prior or dual-stream feature refinement leads to a significant drop in anomaly segmentation performance.

Number of dual-stream feature refinement blocks. Table 4(b) shows the impact of the number of dual-stream feature refinement . While 4 and 6 blocks yield the best AuPRC and FPR95, we choose 4 to minimize trainable parameters.

Anomaly Prior Module	Dual-Stream Feature Refinement	AuPRC↑	FPR_{95}↓
✗	✓	24.9	27.7
✓	✗	68.3	13.1
✓	✓	**89.9**	**1.2**

(a)

Number of Dual-Stream Adapter	AuPRC↑	FPR_{95}↓
1	72.1	5.8
2	87.8	**0.6**
4	**89.9**	1.2
6	81.3	4.7

(b)

Table 4: (a) **Component wise ablation:** We test the importance of each component by removing one at a time. We observe that presence of all components gives the best performance. (b) **Number of dual-stream adapter:** Number of dual-stream adapter as 4 and 6 gives the best AuPRC and FPR_{95}, respectively. All the results in **bold** are best.

6 Conclusion

In this work, we propose a novel adapter for open-set segmentation that disentangles in-distribution and out-of-distribution features via a dual-stream. It consists of an anomaly prior module for initial feature separation and a refinement module leveraging a frozen ViT encoder. Our architecture achieves state-of-the-art results on road obstacle benchmarks while maintaining the best ID performance with fewer trainable parameters than best methods.

Acknowledgment: The authors would like to thank Francesco Pappariello, Principal Engineer, Fabien Castanier, Program Manager, and Viviana D'Alto, Artificial Intelligence Software & Tools Research Platform Director, of STMicroelectronics for their (guidance and) support throughout this work. The would also like to acknowledge that this manuscript reflects only the authors. views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] Jan Ackermann, Christos Sakaridis, and Fisher Yu. Maskomaly: Zero-shot mask anomaly segmentation. In *34th British Machine Vision Conference 2022, BMVC 2022, Aberdeen, UK, November 20-24, 2023*, page 329. BMVA Press, 2023.

[2] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*, 2018.

[3] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *International Journal of Computer Vision*, 129:3119–3135, 2021.

[4] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks,*, 2021.

[5] Robin Chan, Matthias Rottmann, and Hanno 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*, pages 5128–5137, 2021.

[6] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *International Conference on Learning Representations (ICLR)*, 2023.

[7] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021.

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[10] Pau de Jorge, Riccardo Volpi, Puneet K Dokania, Philip HS Torr, and Grégory Rogez. Placing objects in context via inpainting for out-of-distribution segmentation. In *European Conference on Computer Vision*, pages 456–473. Springer, 2024.

[11] Anja Delić, Matej Grcić, and Siniša Šegvić. Outlier detection by ensembling uncertainty with negative objectness. *BMVC*, 2024.

[12] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16918–16927, 2021.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR*, 2021.

[14] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[16] Zhitong Gao, Shipeng Yan, and Xuming He. Atta: Anomaly-aware test-time adaptation for out-of-distribution detection in segmentation. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Matej Grcic and Sinisa Segvic. Hybrid open-set segmentation with synthetic negative data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(10):6748–6760, 2024. doi: 10.1109/TPAMI.2024.3386971. URL https://doi.org/10.1109/TPAMI.2024.3386971.

[18] Matej Grcić, Petra Bevandić, Zoran Kalafatić, and Siniša Šegvić. Dense anomaly detection by robust learning on synthetic negative data. *arXiv preprint arXiv:2112.12833*, 2021.

[19] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Densehybrid: Hybrid anomaly detection for dense open-set recognition. In *European Conference on Computer Vision*, pages 500–517. Springer, 2022.

[20] Matej Grcić, Josip Šarić, and Siniša Šegvić. On advantages of mask-level recognition for outlier-aware segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2936–2946, 2023.

[21] Matej Grcic, Petra Bevandic, Zoran Kalafatic, and Sinisa Segvic. Dense out-of-distribution detection by robust learning on synthetic negative data. *Sensors*, 24(4): 1248, 2024. doi: 10.3390/S24041248. URL https://doi.org/10.3390/s24041248.

[22] Denis Gudovskiy, Tomoyuki Okuno, and Yohei Nakata. Concurrent misclassification and out-of-distribution detection for semantic segmentation via energy-based normalizing flow. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI '23. JMLR.org, 2023.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[25] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, and Jacob Steinhardt. Improving and assessing anomaly detectors for large-scale settings. *OpenReview preprint*, 2021.

[26] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[27] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul 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*, pages 15425–15434, 2021.

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[29] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

[30] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[31] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021.

[32] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[34] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2152–2161, 2019.

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[36] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *CoRR*, 2018.

[37] Nazir Nayal, Misra Yavuz, Joao F Henriques, and Fatma Güney. Rba: Segmenting unknown regions rejected by all. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 711–722, 2023.

[38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

[39] Shyam Nandan Rai, Fabio Cermelli, Barbara Caputo, and Carlo Masone. Mask2anomaly: Mask transformer for universal open-set segmentation. *arXiv preprint arXiv:2309.04573*, 2023.

[40] Shyam Nandan Rai, Fabio Cermelli, Dario Fontanel, Carlo Masone, and Barbara Caputo. Unmasking anomalies in road-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4037–4046, 2023.

[41] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[42] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127, 2018.

[43] Matteo Sodano, Federico Magistri, Lucas Nunes, Jens Behley, and Cyrill Stachniss. Open-World Semantic Segmentation Including Class Similarity . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3184–3194, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.00307. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.00307.

[44] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *European Conference on Computer Vision*, pages 246–263. Springer, 2022.

[45] Yuanpeng Tu, Yuxi Li, Boshen Zhang, Liang Liu, Jiangning Zhang, Yabiao Wang, and Cairong Zhao. Self-supervised likelihood estimation with energy guidance for anomaly segmentation in urban scenes. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 21637–21645. AAAI Press, 2024. doi: 10.1609/AAAI.V38I19.30162. URL https://doi.org/10.1609/aaai.v38i19.30162.

[46] Tomas Vojir, Tomáš Šipka, Rahaf Aljundi, Nikolay Chumerin, Daniel Olmeda Reino, and Jiri Matas. Road anomaly detection by partial image reconstruction with segmentation coupling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15651–15660, 2021.

[47] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *CoRR*, abs/2106.13797, 2021. URL https://arxiv.org/abs/2106.13797.

[48] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023.

[49] Xuanlong Yu, Yi Zuo, Zitao Wang, Xiaowen Zhang, Jiaxuan Zhao, Yuting Yang, Licheng Jiao, Rui Peng, Xinyi Wang, Junpei Zhang, et al. The robust semantic segmentation uncv2023 challenge results. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4620–4630. IEEE Computer Society, 2023.

[50] Dan Zhang, Kaspar Sakmann, William Beluch, Robin Hutmacher, and Yumeng Li. Anomaly-aware semantic segmentation via style-aligned ood augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4065–4073, 2023.

[51] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.