

---

# Generalize or Detect? Towards Robust Semantic Segmentation Under Multiple Distribution Shifts

---

Zhitong Gao<sup>1,2</sup> Bingnan Li<sup>1</sup> Mathieu Salzmann<sup>2</sup> Xuming He<sup>1,3</sup>

<sup>1</sup>ShanghaiTech University <sup>2</sup>EPFL

<sup>3</sup>Shanghai Engineering Research Center of Intelligent Vision and Imaging

{gaozht, libn, hexm}@shanghaitech.edu.cn

mathieu.salzmann@epfl.ch

## Abstract

In open-world scenarios, where both novel classes and domains may exist, an ideal segmentation model should detect anomaly classes for safety and generalize to new domains. However, existing methods often struggle to distinguish between domain-level and semantic-level distribution shifts, leading to poor out-of-distribution (OOD) detection or domain generalization performance. In this work, we aim to equip the model to generalize effectively to covariate-shift regions while precisely identifying semantic-shift regions. To achieve this, we design a novel generative augmentation method to produce coherent images that incorporate both anomaly (or novel) objects and various covariate shifts at both image and object levels. Furthermore, we introduce a training strategy that recalibrates uncertainty specifically for semantic shifts and enhances the feature extractor to align features associated with domain shifts. We validate the effectiveness of our method across benchmarks featuring both semantic and domain shifts. Our method achieves state-of-the-art performance across all benchmarks for both OOD detection and domain generalization. Code is available at <https://github.com/gaozhitong/MultiShiftSeg>.

## 1 Introduction

Semantic segmentation, a fundamental task in computer vision, has become indispensable in various real-world applications, such as autonomous driving [35]. Recent progress in deep learning-based semantic segmentation has exhibited promising results under the assumption of consistent distributions between the training and testing data. However, these models often falter when faced with distributional shifts. Consequently, research on semantic segmentation under distributional shifts has garnered significant attention in recent years. Some studies approach this challenge from a generalization perspective, aiming to train networks to adapt to data with covariate distribution shifts, such as novel domains [9, 45]. Another line of research focuses on training models to discern (or detect) test data exhibiting semantic distributional shifts, such as anomalies or unfamiliar objects, to ensure reliable predictions [5, 4]. In real-world situation, both types of distribution shifts often occur jointly. This leaves us with the question: *Can a model jointly handle both kinds of distribution shift?*

To address this question, we assess the ability of current domain generalization techniques [9, 45] to detect unknown objects and that of out-of-distribution detection techniques [31, 42, 46] to generalize to unknown domains. Interestingly, we find that models trained using domain generalization techniques, such as domain randomization or whitening transformation, often fail to identify unknown objects, and sometimes even perform worse than the baseline without domain generalization. Furthermore, we observe that models trained using out-of-distribution detection techniques struggle to generalize to unknown domains, exhibiting overly high uncertainty towards objects experiencing domain shifts compared to baseline methods without OOD training. While one intuitive approach is

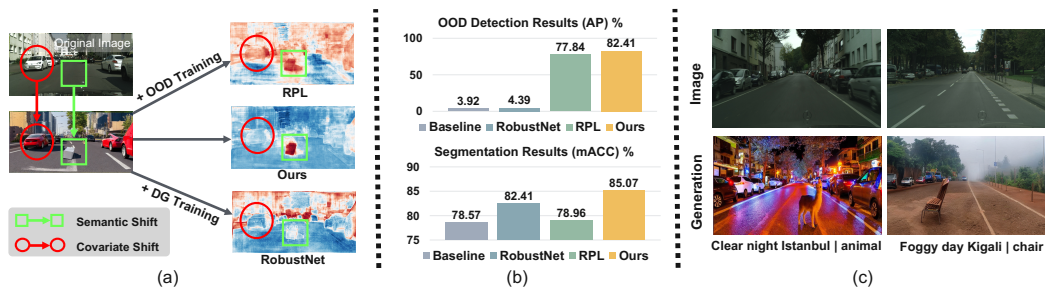


Figure 1: We study semantic segmentation with both **semantic-shift** and **covariate-shift** regions. (a) Training for *Out-of-distribution (OOD) detection* alone [31] yields high uncertainty for both types of shifts, whereas training for *domain generalization (DG)* alone [9] tends to produce low uncertainty for both. Our method effectively differentiates between the two, generating high uncertainty only for semantic-shift regions. (b) We achieve strong performance in both OOD detection and domain-generalized semantic segmentation. (c) This is achieved by coherently augmenting original images (first row) with both covariate and semantic shifts (second row).

to combine existing anomaly segmentation and domain generalization techniques during training, we note that current domain generalization strategies primarily address image-level shifts, whereas anomaly segmentation focuses on object-level semantic differences. Consequently, the resulting models tend to generalize well to image-level variations, such as changes in weather but struggle with object-level shifts. They often misinterpret any object-level distribution shift as a semantic anomaly, assigning high uncertainty scores to known objects that exhibit covariate changes, such as color variations in cars or changes in pedestrian attire, as demonstrated in Fig. 1. These experiments underscore the challenge of differentiating and jointly handling different types of distribution shifts.

In this work, we jointly study both semantic and covariate distribution shifts. That is, we aim to equip the model to generalize effectively to covariate-shift regions while precisely identifying semantic-shift regions. To achieve this, we design a novel generative augmentation method to produce coherent images that incorporate both anomaly (or novel) objects and various covariate shifts at both image and object levels. Furthermore, we introduce a training strategy that re-calibrates uncertainty specifically for semantic shifts and enhances the feature extractor to align features associated with domain shifts<sup>1</sup>.

Specifically, we first introduce a novel data augmentation technique that employs a semantic-to-image generation model to create data that encompasses both covariate and semantic shifts at various levels, allowing the model to learn the essential differences between the shift types. Additionally, we introduce a learnable, semantic-exclusive uncertainty function trained using a relative contrastive loss. We adopt a two-stage training paradigm designed to balance the integration of these enhancements while minimizing their potential interference. A noise-aware training strategy further complements this approach, employing online, pixel-wise selection to mitigate noise in the generated images. Altogether, our approach not only boosts the model's generalization across domain shifts but also ensures a high level of uncertainty in response to semantic shifts.

We validate the effectiveness of our method across benchmarks featuring both semantic and domain shifts, including RoadAnomaly [30], SMIYC [5], ACDC-POC [12] and MUAD [15] benchmarks. Our results demonstrate that our method achieves state-of-the-art performance across all benchmarks, employing different segmentation backbones for both OOD detection and known class segmentation.

In summary, our contributions are: (1) We study semantic segmentation under both semantic and domain shifts, revealing limitations in methods focused on a single shift; (2) We introduce a coherent-generative augmentation method that augments training data with both shifts; (3) We propose a two-stage, noise-aware training pipeline to optimally leverage augmented data, learning a semantic-exclusive uncertainty function while aligning features for domain shifts.

## 2 Related Work

**Anomaly Segmentation** (a.k.a. dense out-of-distribution detection). The task aims to detect anomalies or unknown objects by producing pixel-level uncertainty maps. One approach uses generative

<sup>1</sup>In this work, we use 'domain shift' and 'covariate shift' interchangeably.

models to learn training distributions and detect anomalies through reconstruction differences [30, 48], but this often requires additional networks, resulting in slower inference. Other methods use auxiliary OOD data to train models to distinguish known from unknown instances [46, 6, 31, 42, 36, 18]. Among these, Entropy Maximization [6] uses entire images from COCO [29] as OOD proxies, maximizing softmax entropy on these samples. PEBAL [46] improves upon this by cutting out OOD object instances, pasting them into training images, and using an energy function as the uncertainty score. To reduce artifacts in the pasted OOD region, [52] proposes using a style transfer model to align the pasted region with the background. RPL [31] further regularizes embedding similarity between COCO background pixels and training images. Beyond improvements in OOD proxies and uncertainty functions, recent methods explore the use of the Mask2Former architecture [8], such as RbA [36], Mask2Anomaly [42], and EAM [18]. Our method follows this second approach, generating OOD data with a semantic-to-image model to reduce artifacts and further introducing a learnable uncertainty function to enhance both OOD detection and known class segmentation. It is architecture-agnostic, compatible with both pixel-based and mask-based segmentation backbones.

**Domain Generalization for Semantic Segmentation** The task aims to train a model on one or more source domains that can perform well on unseen target domains. Existing techniques focus either on introducing specialized model architectures, such as those incorporating normalization [39] or whitening transformations [9, 41, 27], or on designing domain randomization techniques [45, 50, 23]. Most domain randomization methods rely on image transformation rules or style transfer [45, 50]. Recently, Jia et al. [23] proposed a semantic-to-image model that generates images across diverse domains. Orthogonally, Bi et al. [3] explore architectural changes with Mask2Former. Our approach belongs to the domain randomization category, generating images with both domain and semantic shifts simultaneously to improve the model’s ability to distinguish between these shifts.

**Segmentation Under Multiple Distribution Shifts** Early works [51, 2] demonstrated the necessity and feasibility of addressing both semantic segmentation under domain shifts and anomaly segmentation. However, these problem settings remain in their early stages (e.g., image-level anomalies) and may not fully capture the true challenges. More recent benchmarks, such as RoadAnomaly [30], SMIYC [5], and MUAD [15], include domain and semantic shifts that better reflect real-world scenarios. Some recent studies [16] have explored the effects of domain shifts on anomaly segmentation benchmarks and proposed a test-time adaptation pipeline to address the problem. In this work, we aim to further bridge the gap by investigating the core challenges of adopting domain generalization techniques and simultaneously enhancing model performance in both areas.

**Generative-based Data Augmentation** This technique is widely used to expand training datasets and prevent overfitting [14, 38, 20, 23, 12]. Unlike rule-based augmentation methods, which focus on image-level changes, generative methods can introduce more object-level variations. Among these works, [12, 33] are the most related to ours, using a text-guided inpainting pipeline to generate anomalies or novel objects. However, this local generation process risks creating inconsistencies between the patch and its background. Additionally, they either focus on generating novel objects within the same domain [12] or use separate pipelines for domain and semantic shifts [33]. In contrast, our method generates multiple distribution shifts in a single process, preserving the global context of the image and ensuring a more natural integration of novel objects.

## 3 Method

### 3.1 Problem Formulation and Method Overview

We consider the problem of *semantic segmentation under multiple distribution shifts*. Formally, we define the training distribution as  $P_{XY}$  in  $\mathcal{X} \times \mathcal{Y}_{in}^{H \times W}$ , where  $\mathcal{X} = \mathbb{R}^{3 \times H \times W}$  represents the three-dimensional input space of images with  $H \times W$  pixels, and  $\mathcal{Y}_{in} = [1, C]$  denotes the semantic label space at each pixel. The test distribution is denoted as  $Q_{XY} \in \mathcal{X} \times \mathcal{Y}_{test}^{H \times W}$ . There are two common types of distribution shifts: *covariate shifts*—where the input distribution changes ( $Q_X \neq P_X$ ) but the label space remains the same—and *semantic shifts*, which involve alterations to the label space, including the introduction of novel categories ( $\mathcal{Y}_{test} \neq \mathcal{Y}_{in}$ ). We consider the possibility of both types of distribution shifts occurring during testing.

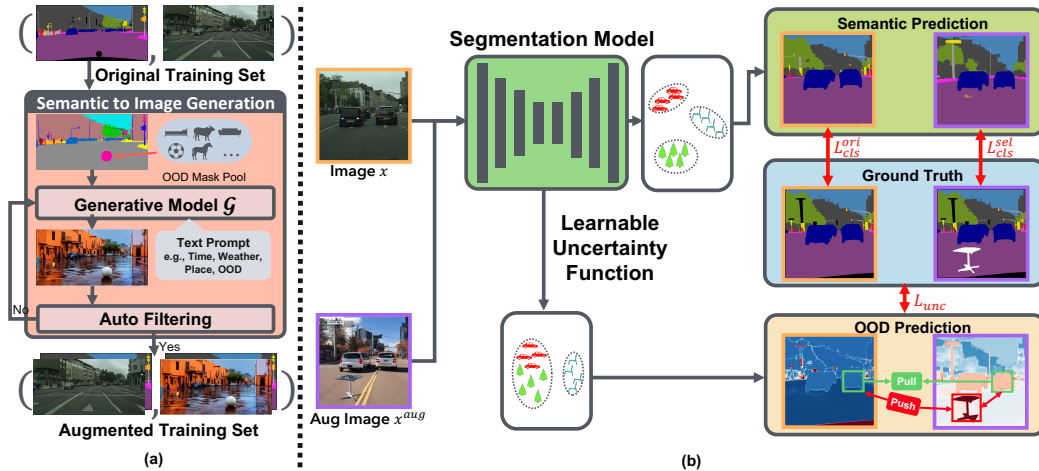


Figure 2: Method Overview: (a) A novel generative-based data augmentation strategy that supplements training data with both covariate and semantic shifts in a coherent manner. (2) A semantic-exclusive uncertainty function with two-stage noise-aware training to encourage invariant feature learning for covariate-shift regions while maintaining high uncertainty for semantic-shift regions.

Our goal is to learn a model capable of jointly identifying semantic-shift regions and generalizing well under covariate shifts. This involves two primary challenges: (1) Enabling the model to distinguish between the two types of distribution shifts, and (2) ensuring the model responds appropriately to each type. To address the first challenge, we introduce a novel generative-based data augmentation strategy that supplements training data with both covariate and semantic shifts in a coherent manner. To tackle the second challenge, we propose a semantic-exclusive uncertainty function with a decoupled training strategy. This encourages the model to learn invariant features for covariate-shift regions while maintaining high uncertainty for semantic-shift regions. Below, we first introduce our generative-based augmentation strategy (Section 3.2), followed by the model training pipeline (Section 3.3).

### 3.2 Coherent Generative-based Augmentation

To distinguish between covariate and semantic shifts, we design a coherent generative-based data augmentation (CG-Aug) pipeline that enriches the training data with realistic and diverse distribution shifts. The pipeline consists of two stages: The first stage uses zero-shot semantic-to-image generation to create a variety of synthetic data, while the second stage automatically filters out low-quality synthetic data. We describe the details of each stage below.

**Zero-Shot Semantic-to-Image Generation.** To generate more realistic and diverse OOD data for segmentation, we propose a generation process that first cut-and-pastes the semantic mask of novel objects to the training labels and then leverages a semantic-to-image generation model to create corresponding augmentation images. By exploiting powerful image generation models, this process is able to produce images with a wide range of covariate shifts and augments the training images with both covariate and semantic shifts in a coherent manner. We detail our process below.

Formally, given training set  $\mathcal{D}^{tr} := \{(x_n, y_n)\}_{n=1}^{N_t}$  with  $(x_n, y_n) \sim P_{XY}$ , we introduce an auxiliary OOD set  $\mathcal{D}^o := \{y_m^o\}_{m=1}^{N_a}$  with object masks  $y_m^o \in \mathcal{Y}_{out}^{H \times W}$ . Subsequently, using a pretrained semantic-to-image generative model  $\mathcal{G} : (\mathcal{Y}_{in} \cup \mathcal{Y}_{out})^{H \times W} \rightarrow \mathbb{R}^{H \times W}$ , we generate an augmented image as

$$x^{aug} = \mathcal{G}(y^{aug}, t) \quad \text{with} \quad y^{aug} = y \oplus y^o, \quad (1)$$

where  $t$  is a text prompt and  $\oplus$  denotes the pasting operation. Here we adopt a pretrained ControlNet [53] as  $\mathcal{G}$  to instantiate the semantic-to-image generation process. Thanks to the powerful prior encoded in Stable Diffusion [43], this process allows us to generate images with more diverse styles than a task-specific semantic-to-image generation model, therefore creating rich covariate shifts. Moreover, we leverage the text prompt  $t$  to produce more diversity in the augmented images by specifying the space, time, and weather, and to enhance the OOD object generation by indicating the class of the pasted objects, via a set of templates (see Appendix A.1 for details).



**Auto-Filtering.** While generative-based augmentation can produce more diverse and realistic distribution shifts than rule-based augmentation, we observed it to often yield inaccurate or noisy rendering for the OOD objects. This might be caused by the fact that these objects appear rarely, or their cut masks are inconsistent with the surroundings. To cope with this, we design an automatic filtering process that identifies generation failures where no object is generated, or a known-category object is incorrectly generated. To achieve this, we leverage pretrained segmentation models to check the region size or its semantic class, and assign a quality score to each generated image. We then filter out the images with low-quality scores (see Appendix A.2 for details).

We perform the above image generation process offline before model training, and the resulting synthetic data is used alongside standard augmentation strategies, such as mixup and AnomalyMix [46], during training. Below, we denote our augmented dataset as  $D^{aug} = \{(x_n, y_n, x_n^{aug}, y_n^{aug})\}_{n=1}^{N_t}$ .

### 3.3 Model Training

Given the augmented dataset, we aim to train a segmentation model with recalibrated uncertainty output, generating high OOD scores for semantic-shift regions and performing robustly under covariate shifts. To achieve this, we propose a learnable uncertainty function and develop a stage-wise learning strategy that initializes the uncertainty function before fine-tuning the entire model. Our training process integrates a relative contrastive loss and a noise-aware data selection scheme, enabling the model to effectively align both the feature space and the OOD output scores. We note that our method is generic and can be applied to pixel-wise models (e.g. DeepLabv3+ [7]) or mask-wise models (e.g. Mask2Former [8]).

#### 3.3.1 Semantic-Exclusive Uncertainty Recalibration

**Learnable Uncertainty Function.** Suppose we have a neural network with its feature extractor  $f(x) \in \mathbb{R}^{M \times F}$ , where  $M$  is the number of pixels (or masks), and  $F$  is the feature dimension. We introduce a learnable linear projection  $W^o \in \mathbb{R}^{F \times C}$ , with  $W_c^o$  denotes  $W^o[:, c]$  for short. For a pixel-wise prediction model, we adopt the energy function form and parameterize it into a learnable uncertainty function

$$u(x) = \log \sum_c \exp f(x) W_c^o. \quad (2)$$

For a mask-based segmentation network, we use the adapted maximum softmax probability (MSP) defined in [8] as the uncertainty function, and parameterize it with  $W^o$ , leading to

$$u(x) = \max_c \left( \text{softmax}(f(x) W_c^o)^T \cdot g(x) \right). \quad (3)$$

Here  $g(x) \in (0, 1)^{M \times H \times W}$  is the sigmoid output of the mask head. For both cases, we initialize the projection function  $W^o$  as the class weight  $W^{in}$  of the pretrained segmentation network. The corresponding uncertainty score corresponds to the original energy score (or MSP score).

**Relative Contrastive Loss.** We train the uncertainty function using a novel relative contrastive loss, which encourages higher uncertainty in unknown-class regions compared to known-class regions, while ensuring that regions with and without covariate shifts exhibit similar levels of uncertainty.

Formally, for each batch of data  $\{(x_n, y_n, x_n^{aug}, y_n^{aug})\}_{n=1}^B$ , where  $B$  is the batch size, we define the following pixel index sets:  $\Omega^{in} = \{i : y(i) \in \mathcal{Y}_{in}\}$ , representing *inlier* pixel indices from the original training images;  $\Omega^{aug} = \{i : y^{aug}(i) \in \mathcal{Y}_{in}\}$ , representing *inlier* pixel indices from the *augmented* training images (covariate-shift set); and  $\Omega^{out} = \{i : y(i) \notin \mathcal{Y}_{in}\} \cup \{i : y^{aug}(i) \notin \mathcal{Y}_{in}\}$ , representing *outlier* pixel indices from both original and augmented images (semantic-shift set). Here,  $y(i)$  (or  $y^{aug}(i)$ ) denotes the label of pixel  $i$ . Our contrastive loss is defined as

$$L_{unc} = \sum_{o \in \Omega^{out}, i \in \Omega^{in}} \tau_{\lambda_1}(u_o - u_i) + \sum_{o \in \Omega^{out}, c \in \Omega^{aug}} \tau_{\lambda_2}(u_o - u_c) + \sum_{c \in \Omega^{aug}, i \in \Omega^{in}} m_{c,i} \cdot \tau_{\lambda_3}(-(u_c - u_i)), \quad (4)$$

where  $\tau_\lambda(x) = \max(\lambda - x, 0)$  is the margin-based contrastive loss, which encourages the input value to exceed  $\lambda$ . The first two terms promote larger uncertainty gaps between unknown-class and known-class regions, while the third term encourages smaller uncertainty gaps between covariate-shifted and

original data. For the third term, we calculate gaps only between pairs of original and augmented images, with  $m_{c,i} \in \{0, 1\}$  indicating whether pixel  $(c, i)$  is paired in the dataset. The three margin values  $(\lambda_1, \lambda_2, \lambda_3)$  introduce priors on the uncertainty gaps, and are set based on the initial average distance. Our method remains robust across a wide range of margin values (cf. Table 6).

Compared to existing OOD losses that either maximize uncertainty only for unknown data [31] or supervise known and unknown data separately [46, 42], our loss supervises the relative distance between them, making it more robust to hyperparameters and simpler to train (cf. Sec.4.5).

### 3.3.2 Two-Stage Noise-Aware Training

We now present our two-stage training procedure, which sequentially learns the uncertainty function and the feature extractor  $f(x)$  of the segmentation network. Specifically, we first freeze the pre-trained segmentation network and learn the semantic-exclusive uncertainty function using the relative contrastive loss defined in Eq. 4. We then fine-tune the feature extractor with both the contrastive and standard segmentation loss to improve the feature representations of both known and OOD classes.

Despite the offline filtering process, the generated images may still contain regions that are inconsistent with the label masks. To address this, we introduce a pixel-wise sample selection scheme during training, based on the ‘small loss’ criterion [1]. Specifically, we compute and rank the cross-entropy loss for each pixel, selecting pixels with smaller losses for backpropagation while ignoring those with larger losses. Formally, our *selective cross-entropy loss* is defined as

$$L_{\text{seg}}(y, p, \eta) = \sum_i \eta_i \sum_c y_i^c \log p_i^c, \quad (5)$$

where  $p_i$  and  $y_i$  represent the pixel-wise softmax score and one-hot label, respectively, and  $\eta_i \in \{0, 1\}$  indicates whether a pixel is selected for backpropagation. We determine the percentage of selected pixels per batch by visualizing the selection map of a small number of samples, ensuring that visibly incorrect patterns are excluded (see Figure 4 for an example). For models using the Dice loss, such as mask-based ones, we use a similar scheme to remove pixels with a large loss (cf. Appendix A.3).

For the original data, which we assume to be noise-free, we set  $\eta_i = 1$  for all pixels. This corresponds to using the *standard cross-entropy loss*. We denote the segmentation loss for the original data as  $L_{\text{seg}}^{\text{in}}$  and for the generated augmentation data as  $L_{\text{seg}}^{\text{aug}}$ . The overall loss function can be written as

$$L = L_{\text{unc}} + \beta_1 L_{\text{seg}}^{\text{in}} + \beta_2 L_{\text{seg}}^{\text{aug}}. \quad (6)$$

Here,  $\beta_1$  and  $\beta_2$  ensure that the three loss terms are on the same scale.

In summary, our semantic-exclusive uncertainty function, trained through a decoupled parameter training approach and relative contrastive loss, enables the model to fully leverage the generated distribution-shift data. Our noise-aware learning strategy enhances the model’s robustness against generation errors. Together, these components of our training pipeline equip the model to effectively learn both domain generalization and accurate OOD detection, ensuring robust performance in dynamic open-world scenarios.

## 4 Experiments

In this section, we evaluate our method’s performance in jointly handling anomaly segmentation and domain generalization using several datasets that include both domain and semantic shifts: RoadAnomaly [30], SMIYC [5], ACDC-POC [12], and MUAD [15]. We first introduce the datasets in Sec.4.1 and describe the experimental setup in Sec.4.2. The results are presented in Sec.4.3 and Sec.4.4, followed by an ablation study in Sec. 4.5.

### 4.1 Datasets

Following the literature [5, 31, 8], we train our model on the Cityscapes dataset [11] and evaluate its performance on the test sets described below. Based on the evaluation goals, we divide these datasets into two groups. Examples from each dataset are shown in Fig. 3.

**Anomaly Segmentation Datasets:** (a) *The Road Anomaly dataset* [30] includes 60 images of real-world road anomalies such as animals, rocks, and obstacles, featuring various driving conditions

Table 1: **Results on anomaly segmentation benchmarks:** RoadAnomaly, SMIYC-RA21 and SMIYC-RO21. Our method achieves the best results under both backbones (Best results in Bold).

| Method                 | Backbone    | RoadAnomaly    |               |                                | SMIYC - RA21  |                                | SMIYC - RO21  |                                |
|------------------------|-------------|----------------|---------------|--------------------------------|---------------|--------------------------------|---------------|--------------------------------|
|                        |             | AUC $\uparrow$ | AP $\uparrow$ | FPR <sub>95</sub> $\downarrow$ | AP $\uparrow$ | FPR <sub>95</sub> $\downarrow$ | AP $\uparrow$ | FPR <sub>95</sub> $\downarrow$ |
| Maximum softmax [21]   | DeepLabv3+  | 67.53          | 15.72         | 71.38                          | 27.97         | 72.05                          | 15.72         | 16.60                          |
| ODIN [28]              |             | -              | -             | -                              | 33.06         | 71.68                          | 22.12         | 15.28                          |
| Mahalanobis [26]       |             | 62.85          | 14.37         | 81.09                          | 20.04         | 86.99                          | 20.90         | 13.08                          |
| Image resynthesis [30] |             | -              | -             | -                              | 52.28         | 25.93                          | 37.71         | 4.70                           |
| SynBoost [13]          |             | 81.91          | 38.21         | 64.75                          | 56.44         | 61.86                          | 71.34         | 3.15                           |
| Maximized entropy [6]  |             | -              | 48.85         | 31.77                          | 85.47         | 15.00                          | 85.07         | 0.75                           |
| PEBAL [46]             |             | 87.63          | 45.10         | 44.58                          | 49.14         | 40.82                          | 4.98          | 12.68                          |
| Dense Hybrid [17]      |             | -              | 31.39         | 63.97                          | 77.96         | 9.81                           | 87.08         | <b>0.24</b>                    |
| RPL+CoroCL [31]        |             | 95.72          | 71.61         | 17.74                          | 83.49         | 11.68                          | 85.93         | 0.58                           |
| Ours                   |             | <b>96.40</b>   | <b>74.60</b>  | <b>16.08</b>                   | <b>88.06</b>  | <b>8.21</b>                    | <b>90.71</b>  | 0.26                           |
| Mask2Anomaly [42]      | Mask2Former | -              | 79.70         | 13.45                          | 88.7          | 14.60                          | 93.3          | 0.20                           |
| RbA [36]               |             | -              | 85.42         | 6.92                           | 90.90         | 11.60                          | 91.80         | 0.50                           |
| M2F-EAM [18]           |             | -              | 69.40         | 7.70                           | <b>93.75</b>  | <b>4.09</b>                    | 92.87         | 0.52                           |
| Ours                   |             | <b>97.94</b>   | <b>90.17</b>  | <b>7.54</b>                    | 91.92         | 7.94                           | <b>95.29</b>  | <b>0.07</b>                    |

and covariate shifts. (b) *The SMIYC benchmark* [5] consists of RoadAnomaly21 (10 validation, 100 test images) and RoadObstacle21 (30 validation, 327 test images), with anomaly objects and domain shifts. These datasets provide masks for anomaly objects, allowing us to evaluate our method’s performance on anomaly segmentation under distribution shifts.

**Joint Anomaly Segmentation and Domain Generalization Datasets:** (a) *The ACDC-POC dataset* [12] is based on the original ACDC Validation set [44] with generated anomaly objects via inpainting [12]. It contains 200 images with domain shifts including various weather and night scenes. (b) *The MUAD dataset* [15] is a synthetic dataset containing various driving environments and anomaly objects. We use the challenge test set as in [49], which contains 240 images with domain shifts at both object and image levels, and anomaly objects such as animals and trash cans.<sup>2</sup> These two datasets contain both known-class annotations and unknown object masks, enabling us to evaluate our method jointly for anomaly segmentation and domain generalization.

## 4.2 Experimental Setup

**Performance Measure:** For evaluation of anomaly segmentation, we use the Area Under the Receiver Operating Characteristics curve (AUROC), the Average Precision (AP), and the False Positive Rate at a True Positive Rate of 95% (FPR95). For evaluation of known class segmentation, we use the mean intersection-over-union (mIoU) and the mean accuracy (mAcc).

**Implementation Details:** We build our method on two segmentation backbones: (a) DeepLabv3+ [7] and (b) Mask2Former [8]. We maintain the network architecture, pretrained models, segmentation loss, and training pipeline the same as in previous work [31, 42] to make a fair comparison. We use the SMIYC validation set for model selection and maintain the same model for evaluation across all test sets. We refer the reader to Appendix A for other training details.

## 4.3 Results on Anomaly Segmentation Benchmarks

We present the performance of our method on anomaly segmentation benchmarks, including RoadAnomaly and SMIYC (RA21 and RO21). As shown in Table 1, our method achieves state-of-the-art performance on both DeepLabv3+ and Mask2Former-based models. With the same backbone, it outperforms RPL [31] by 3% on RoadAnomaly and 5% on SMIYC, and surpasses Mask2Anomaly [42] by 10% on RoadAnomaly and 3% on SMIYC. Recent methods, M2F-EAM [18] and RbA [36], use a more powerful Swin Transformer backbone, while ours uses ResNet-50, as Mask2Anomaly. M2F-EAM also uses Mapillary Vistas [37] as additional dataset for training. Despite these unfair comparisons, our method still outperforms both on most metrics, demonstrating its effectiveness.

<sup>2</sup>Since we use Cityscapes as the training set, the unknown object set differs from that used in [49].

Table 2: **Results on ACDC-POC and MUAD.** Our model achieves the best performance in both anomaly segmentation ( $AP\uparrow$ ,  $FPR\downarrow$ ) and domain-generalized segmentation ( $mIoU\uparrow$ ,  $mAcc\uparrow$ ). Anomaly segmentation methods typically perform worse than the baseline for known class segmentation, while domain generalization methods fall below the baseline on OOD detection. (Best results are in bold; results below baseline are in blue.)

| Method             | Backbone    | Technique |    | ACDC-POC     |                      |                |                | MUAD         |                      |                |                |
|--------------------|-------------|-----------|----|--------------|----------------------|----------------|----------------|--------------|----------------------|----------------|----------------|
|                    |             | OOD       | DG | $AP\uparrow$ | $FPR_{95}\downarrow$ | $mIoU\uparrow$ | $mAcc\uparrow$ | $AP\uparrow$ | $FPR_{95}\downarrow$ | $mIoU\uparrow$ | $mAcc\uparrow$ |
| Baseline [7]       | DeepLabv3+  | -         | -  | 3.92         | 55.50                | 46.89          | 78.57          | 1.34         | 72.78                | 29.47          | 68.63          |
| RuleAug [45]       |             | -         | ✓  | 2.09         | 72.79                | 48.60          | 81.79          | 0.99         | 81.08                | 29.42          | 69.22          |
| RobustNet [9]      |             | -         | ✓  | 4.39         | 62.65                | 47.41          | 82.41          | 2.27         | 58.64                | <b>32.18</b>   | 72.02          |
| PEBAL [46]         |             | ✓         | -  | 20.67        | 14.35                | 45.59          | 81.28          | 7.81         | 47.56                | 29.08          | 66.41          |
| RPL [31]           |             | ✓         | ✓  | 77.84        | 1.20                 | 46.35          | 78.96          | 27.70        | 24.45                | 29.86          | 71.60          |
| OOD + RuleAug [45] |             | ✓         | ✓  | 80.65        | 1.30                 | 46.76          | 73.08          | 20.97        | 20.37                | 27.83          | 63.02          |
| Ours               |             | ✓         | ✓  | <b>82.41</b> | <b>1.01</b>          | <b>54.12</b>   | <b>85.07</b>   | <b>36.08</b> | <b>18.74</b>         | 31.33          | <b>73.13</b>   |
| Mask2Anomaly [42]  | Mask2Former | ✓         | -  | 73.77        | 3.60                 | 47.32          | 83.10          | 39.32        | 41.24                | 23.43          | 61.91          |
| OOD + RuleAug [45] |             | ✓         | ✓  | 82.82        | 0.79                 | 50.36          | 82.83          | 25.43        | 41.15                | 26.27          | 67.51          |
| Ours               |             | ✓         | ✓  | <b>90.42</b> | <b>0.46</b>          | <b>51.75</b>   | <b>83.16</b>   | <b>45.65</b> | <b>24.70</b>         | <b>28.44</b>   | <b>73.77</b>   |

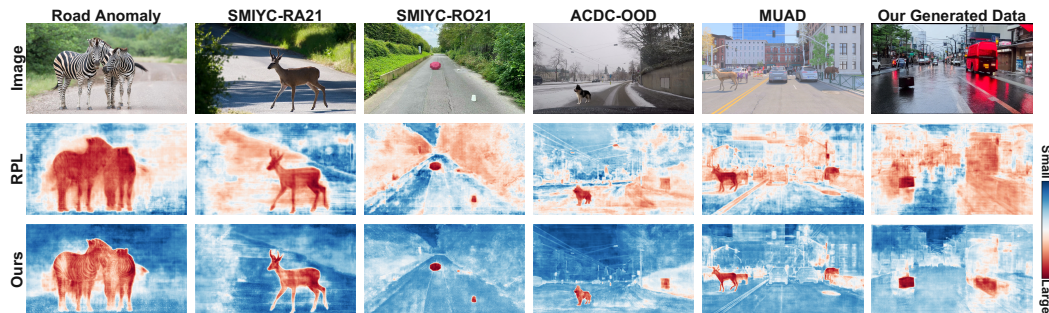


Figure 3: **Comparison of Uncertainty Maps.** Our method robustly detects anomalies under covariate shifts across five datasets (first five columns) and generated data (last column). The previous method RPL [31] failed to distinguish domain from semantic shifts, producing high uncertainty in both cases.

In Fig. 3, we visualize the uncertainty map output by our method using the DeepLabv3+ architecture. Compared to the previous state-of-the-art method, RPL [31], our model assigns higher uncertainty scores to anomalous objects and lower uncertainty scores to covariate shifts. This highlights the efficacy of our method in distinguishing between domain shifts and semantic shifts.

#### 4.4 Results on ACDC-POC and MUAD

We then extend our evaluation to the ACDC-POC and MUAD datasets, assessing both anomaly segmentation performance and known-class domain generalization performance. For a comprehensive comparison, we include both previous state-of-the-art OOD detection techniques and domain generalization techniques [9, 45]. Additionally, we trained a DG+OOD combination method by combining naive OOD training with contrastive loss and rule-based data augmentation (denoted as OOD+RuleAug). A DeepLabv3+ model with standard training is used as a baseline method.<sup>3</sup>

The results are shown in Table 2, where our model achieves the best results for both out-of-distribution detection and domain generalization, demonstrating its capacity in jointly handling both types of distribution shifts. By comparison, previous methods fall short in either known class segmentation or OOD detection. Specifically, we find that: (a) Previous works that mainly focus on **domain generalization** (RobustNet [9], RuleAug [45]) generally improve the known class segmentation results, but their performance in OOD detection is affected, sometimes worse than the baseline. (b) Previous works that mainly focus on **OOD detection** (such as PEBAL [46]) perform poorly

<sup>3</sup>For all compared methods, except RuleAug, we used the official pretrained models provided by the respective authors and performed re-inference to obtain the results. For RuleAug, we applied a combination of color jittering, Gaussian blur, etc., as suggested in [45]. For further details, please refer to the Appendix A.5

Table 3: **Impact of CG-Aug and Training Strategy.** The proposed coherent generative-based augmentation consistently enhances the previous OOD method, Mask2Anomaly [42] (M2A for short). Our fine-tuning strategy makes better use of the data and further boosts the performance.

|          |         | RoadAnomaly  |                     | SMIYC-RA Val |                     | SMIYC-RO Val |                     |
|----------|---------|--------------|---------------------|--------------|---------------------|--------------|---------------------|
| Training | Aug.    | AP↑          | FPR <sub>95</sub> ↓ | AP↑          | FPR <sub>95</sub> ↓ | AP↑          | FPR <sub>95</sub> ↓ |
| M2A [42] | Default | 79.70        | 13.45               | 94.50        | 3.30                | 88.60        | 0.30                |
| M2A [42] | Ours    | 85.47        | 22.38               | <b>97.96</b> | 1.55                | 89.80        | <b>0.12</b>         |
| Ours     | Ours    | <b>90.17</b> | <b>7.54</b>         | 97.31        | <b>1.04</b>         | <b>93.24</b> | 0.14                |

on domain generalization, sometimes worse than the baseline. Furthermore, their OOD detection performance may also be affected by the domain shift. (c) Previous works that jointly handle **image-level DG and OOD** (RPL [31] and OOD+RuleAug) may not fully distinguish object-level domain shifts. Our method leveraging diverse augmentations and a dedicated training strategy enables the model to jointly handle OOD detection and domain generalization.

In Appendix C.1, we provide additional results on individual domain shifts (fog, rain, snow, night) and per-class evaluation. Furthermore, we compare our method with other DG methods on the original ACDC dataset in Appendix C.4, where we show superior domain generalization performance.

#### 4.5 Analysis and Ablation Study

We conduct ablation studies to evaluate the design of our components. We begin by analyzing the effectiveness of our proposed modules: the coherent generative-based augmentation (CG-Aug) and our model training strategy. We then proceed with a detailed examination of each module’s design.

**Impact of CG-Aug and Training Strategy** We evaluate the decoupled contributions of our data augmentation and training strategies in Table 3. Starting with a recent anomaly segmentation method, Mask2Anomaly [42], we first replace its original OOD data, which utilizes cut-and-pasted COCO images, with our proposed CG-Aug. As shown in Row #2, this substitution results in consistent performance improvements across all datasets. This demonstrates the efficacy of introducing data with both semantic and domain shifts in a coherent way. Next, we replace their training strategy with ours, leading to further gains in performance. This indicates that our training strategy is more effective in leveraging the generated data. Additionally, we present and discuss similar experiments using RPL [31] as a baseline. For more details, please refer to Appendix Table 7.

**Ablation Study of our CG-Aug** The proposed CG-Aug generate semantic-shift and domain-shift jointly in a coherent way. To evaluate the design, we compare with three variations: (1) *Semantic-Shift Only (SS)*: Generate images with semantic shift using POC [12]. (2) *Domain-shift or Semantic-shift (DS or SS)*: Create a mixed dataset with either domain shifts (DS) using our semantic-mask-to-image process or semantic shifts (SS) using POC. (3) *Domain-shift and Semantic-shift (DS and SS)*: First generate DS data, then inpaint unknown objects. The second and third methods can be seen as applying [33] to our problem in two ways. Results in Table 4 show that: adding domain shift data significantly improves performance over semantic-shift-only data. Jointly generating DS and SS in one image yields better results than generating them separately. Our method, which generates both DS and SS in one step, achieves the best performance, ensuring more coherence without artifacts and outperforming the two-step approach. We include more comparison results with POC [12] in Appendix C.3.

Table 4: **Ablation Study of CG-Aug.** Generating data with both Semantic-shift (SS) and Domain-shift (DS) in a coherent manner achieves better results than other variations. The experiments were conducted using the Mask2Former backbone and evaluated on the RoadAnomaly dataset.

|               | AUC↑         | AP↑          | FPR <sub>95</sub> ↓ |
|---------------|--------------|--------------|---------------------|
| POC [12] (SS) | 95.43        | 83.66        | 10.33               |
| DS or SS      | 95.90        | 87.64        | 9.28                |
| DS and SS     | 96.47        | 89.08        | 8.16                |
| CG-Aug (Ours) | <b>97.94</b> | <b>90.17</b> | <b>7.54</b>         |

Table 5: **Ablation Study of Our Training Pipeline:** Learnable Uncertainty Function (Learnable-UF), Relative Contrastive Loss (RelConLoss), and Noise-aware Sample Selection (Selection). Experiments are conducted under DeepLabv3+ architecture.

| Learnable-UF | RelConLoss   | Selection    | SMIYC-RA Val  |                       | SMIYC-RO Val  |                       | MUAD          |                       |                 |                 | ACDC -POC     |                       |                 |                 |
|--------------|--------------|--------------|---------------|-----------------------|---------------|-----------------------|---------------|-----------------------|-----------------|-----------------|---------------|-----------------------|-----------------|-----------------|
|              |              |              | AP $\uparrow$ | FPR $_{95}\downarrow$ | AP $\uparrow$ | FPR $_{95}\downarrow$ | AP $\uparrow$ | FPR $_{95}\downarrow$ | mIoU $\uparrow$ | mAcc $\uparrow$ | AP $\uparrow$ | FPR $_{95}\downarrow$ | mIoU $\uparrow$ | mAcc $\uparrow$ |
| $\times$     | $\checkmark$ | $\checkmark$ | 89.34         | 7.51                  | 94.96         | 0.21                  | 24.10         | 25.73                 | 31.67           | 72.16           | 77.92         | 2.00                  | 53.43           | 84.46           |
| $\checkmark$ | $\times$     | $\checkmark$ | 91.01         | 5.78                  | 94.95         | 0.20                  | 20.49         | 24.58                 | 32.68           | <b>73.86</b>    | 75.67         | 1.60                  | 53.22           | 84.32           |
| $\checkmark$ | $\checkmark$ | $\times$     | 91.64         | 4.18                  | <b>96.07</b>  | <b>0.15</b>           | 20.24         | 22.57                 | 31.73           | 71.88           | 77.99         | 1.27                  | 54.26           | <b>85.30</b>    |
| $\checkmark$ | $\checkmark$ | $\checkmark$ | <b>93.82</b>  | <b>3.94</b>           | 95.20         | 0.19                  | <b>36.08</b>  | <b>18.74</b>          | <b>31.33</b>    | 73.13           | <b>82.41</b>  | <b>1.01</b>           | <b>54.12</b>    | 85.07           |

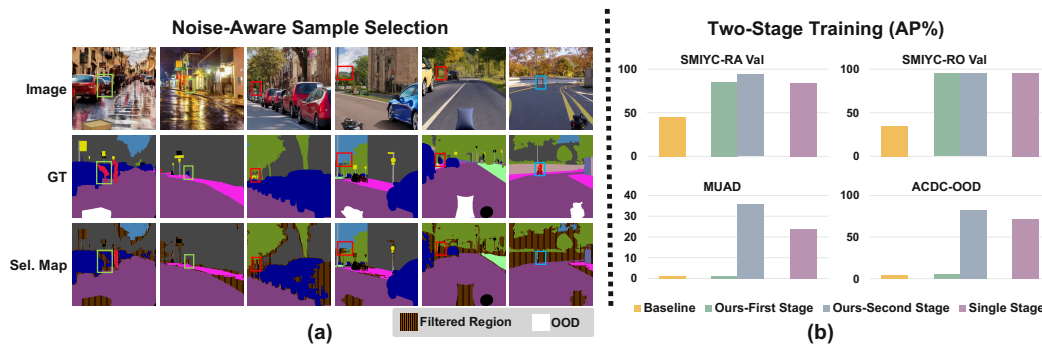


Figure 4: (a) **Visualization of Our Selection Maps.** Our selection strategy effectively identifies and removes generation errors (highlighted with boxes). (b) **Analysis of Our Two-Stage Training.** The first stage of training the uncertainty function boosts baseline performance, and second-stage fine-tuning further improves performance, achieving better results than single-stage training.

**Ablation study of our Training Design** We evaluate our training design in Table 5. In Row #1, we replace our *Learnable Uncertainty Function* (Learnable-UF) with a fixed energy function. In Row #2, we substitute our *Relative Contrastive Loss* (RelConLoss) with an absolute contrastive loss [46, 31, 42], which directly supervises the uncertainty score value rather than the relative gap between two uncertainty scores. In Row #3, we remove the *Sample Selection* module. Compared to our complete method, presented in the final row, these modifications result in decreased performance in both OOD detection and domain generalization, highlighting the effectiveness of our module design. A visualization of the sample selection process is shown in Figure 4 (a).

In Figure 4(b), we evaluate the effectiveness of our *Stage-wise Training* pipeline. Starting from a pre-trained baseline model, our first stage—fine-tuning only the learnable uncertainty function—doubles the performance on SMIYC-RA/RO datasets, demonstrating that the initial uncertainty function is often sub-optimal and can be significantly improved using fixed features [24]. A second-stage feature fine-tuning further boosts performance. Additionally, our two-stage approach outperforms single-stage fine-tuning with a learnable uncertainty function, showing that training directly with uncalibrated uncertainties can disrupt feature learning and degrade OOD detection performance.

We also demonstrate the robustness of our method under a range of hyperparameters (loss margins and selection ratio) in Appendix B.1, and evaluate the impact of generated dataset size in Appendix B.2.

## 5 Conclusion

In this work, we have studied semantic segmentation under multiple distribution shifts, finding that prior methods focusing separately on domain generalization and anomaly segmentation may not effectively handle these complex shifts. To tackle this, we have introduced a coherent generative data augmentation approach that enriches training data with both domain and semantic shifts. Additionally, we have proposed a learnable uncertainty function, trained in a stage-wise manner, to fully utilize the data and produce uncertainty scores specifically for semantic shifts. One limitation of our method is its reliance on the quality of the generative model. While we mitigate generation failures through offline autofiltering and online sample selection, some impact remains, such as lower performance for classes the generative model struggles with and potential limitations in scaling up the generated data (see Appendix E for details).

## Acknowledgement

This work was supported by NSFC 62350610269, Shanghai Frontiers Science Center of Human-centered Artificial Intelligence, and MoE Key Lab of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University).

## References

- [1] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*. PMLR, 2017.
- [2] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Simultaneous semantic segmentation and outlier detection in presence of domain shift. In *Pattern Recognition: 41st DAGM German Conference, DAGM GCPR 2019, Dortmund, Germany, September 10–13, 2019, Proceedings 41*. Springer, 2019.
- [3] Qi Bi, Shaodi You, and Theo Gevers. Learning content-enhanced mask transformer for domain generalized urban-scene segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 819–827, 2024.
- [4] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *arXiv preprint arXiv:1904.03215*, 2019.
- [5] 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. *arXiv preprint arXiv:2104.14812*, 2021.
- [6] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. *arXiv preprint arXiv:2012.06575*, 2020.
- [7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.
- [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*, 2022.
- [9] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T. Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11580–11590, June 2021.
- [10] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11580–11590, 2021.
- [11] 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 *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3213–3223, 2016.
- [12] Pau de Jorge, Riccardo Volpi, Puneet K Dokania, Philip HS Torr, and Gregory Rogez. Placing objects in context via inpainting for out-of-distribution segmentation. *arXiv preprint arXiv:2402.16392*, 2024.



- [13] 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.
- [14] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *Advances in neural information processing systems*, 36:79024–79034, 2023.
- [15] Gianni Franchi, Xuanlong Yu, Andrei Bursuc, Angel Tena, Rémi Kazmierczak, Séverine Dubuisson, Emanuel Aldea, and David Filliat. Muad: Multiple uncertainties for autonomous driving, a benchmark for multiple uncertainty types and tasks. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.
- [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*, 2023.
- [17] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Densehybrid: Hybrid anomaly detection for dense open-set recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, pages 500–517. Springer, 2022.
- [18] 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 2937–2947, 2023.
- [19] 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.
- [20] Sobhan Hemati, Mahdi Beitollahi, Amir Hossein Estiri, Bassel Al Omari, Xi Chen, and Guojun Zhang. Cross domain generative augmentation: Domain generalization with latent diffusion models. *arXiv preprint arXiv:2312.05387*, 2023.
- [21] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
- [22] Lei Huang, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4874–4883, 2019.
- [23] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *European Conference on Computer Vision*, pages 91–109. Springer, 2025.
- [24] 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*, 2021.
- [25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [26] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 7167–7177, 2018.
- [27] Yumeng Li, Dan Zhang, Margret Keuper, and Anna Khoreva. Intra-source style augmentation for improved domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 509–519, 2023.
- [28] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2018.

- [29] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.
- [30] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *Proc. of IEEE international conference on computer vision (ICCV)*, 2019.
- [31] Yuyuan Liu, Choubo Ding, Yu Tian, Guansong Pang, Vasileios Belagiannis, Ian Reid, and Gustavo 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.
- [32] 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.
- [33] Thibaut Loiseau, Tuan-Hung Vu, Mickael Chen, Patrick Pérez, and Matthieu Cord. Reliability in semantic segmentation: Can we use synthetic data? *arXiv preprint arXiv:2312.09231*, 2023.
- [34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [35] S Minaee, YY Boykov, F Porikli, AJ Plaza, N Kehtarnavaz, and D Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [36] 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.
- [37] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Buló, and Peter Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, 2017.
- [38] Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Gustavo Adolfo Vargas Hakim, David Osowiecki, Moslem Yazdanpanah, Ismail Ben Ayed, and Christian Desrosiers. Feedback-guided domain synthesis with multi-source conditional diffusion models for domain generalization. *arXiv preprint arXiv:2407.03588*, 2024.
- [39] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the european conference on computer vision (ECCV)*, 2018.
- [40] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1863–1871, 2019.
- [41] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.
- [42] 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*, 2023.
- [43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.
- [44] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

- [45] Manuel Schwonberg, Fadoua El Bouazati, Nico M Schmidt, and Hanno Gottschalk. Augmentation-based domain generalization for semantic segmentation. In *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023.
- [46] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 246–263. Springer, 2022.
- [47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [48] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan L Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 2020.
- [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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [50] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2100–2110, 2019.
- [51] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. Wilddash-creating hazard-aware benchmarks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–416, 2018.
- [52] Dan Zhang, Kaspar Sakmann, William Beluch, Robin Huttmacher, and Yumeng Li. Anomaly-aware semantic segmentation via style-aligned ood augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [53] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [54] 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.
- [55] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

## A Implementation Details

### A.1 Zero-Shot Semantic-to-Image Generation

We adopt a pretrained semantic-to-image generation model, ControlNet 1.0 [53], provided by the official GitHub repository<sup>4</sup>, for our generation process. This model is based on Stable Diffusion [43] and fine-tuned on ADE20K [54, 55]. It takes two main inputs: a semantic mask and a text prompt.

To obtain our pasted semantic mask, we first convert the masks from Cityscapes labels to ADE20K labels, then overlay this with auxiliary out-of-distribution (OOD) object masks. Specifically, we use the mask labels from ADE20K that belong to the 'thing' categories, excluding those with labels shared with Cityscapes.

Our text prompts have two parts: one part specifies the domain shifts, and the other specifies the OOD objects. For domain-shift prompts, we use the template "An image sampled from various stereo video sequences taken by dash cam in {PLACE} in a {WEATHER} {TIME}", where we define PLACE as a set of 100 cities worldwide, WEATHER as ['cloudy', 'rainy', 'snowy', 'foggy', 'clear'], and TIME as ['day', 'night']. Additionally, we improve the OOD object generation by indicating the specific class of pasted objects in the prompt with the template: "There is a {OOD} accidentally staying on the road.", where OOD is the class name of the pasted object. This further contextualizes the generated scene to reflect realistic anomaly scenarios.

### A.2 Auto-Filtering of Failed Generations

The generation process can be noisy, particularly when generating image regions for pasted OOD object masks. By nature, these objects are anomalies within the scenes, appearing rarely, and their cut masks may exhibit shapes or poses inconsistent with their surroundings. As a result, the generated objects may deviate significantly in shape from the intended mask, be overlooked (blending into the surroundings), or be incorrectly generated as more common objects within the scene. Such discrepancies make the raw augmented image-label pairs too noisy for direct training.

To address these issues, we design an automatic filtering process to identify generation failures, such as cases where no object is generated or a known-category object is incorrectly produced. If the generated object is present and does not belong to a known category, we retain the image while revising the corresponding mask for the generated novel object. Otherwise, we discard the image and regenerate it. To implement this, we use the Segment-Anything Model (SAM), providing the bounding box location of the pasted mask as input to obtain a segmentation. We then compare the SAM output with the original mask, identifying it as a failure case if the Intersection over Union (IoU) is very low (below 0.7). Additionally, we employ a pre-trained segmentation model to produce an uncertainty score for the generated objects, filtering out those with very low uncertainty scores, as these are likely to have been misgenerated into a known-category object. This comprehensive filtering process effectively enhances the quality of the training dataset, making it better suited for effective model training.

### A.3 Training Details on Mask2Former Backbone

Following Mask2Anomaly [42], we train Mask2Former [8] using a combination of dice loss and binary cross-entropy (BCE) loss for the mask prediction head, and cross-entropy loss for the class prediction head. For the dice and BCE losses, we modify the sampling strategy for generated images to implement the sample selection process described in Sec. 3.3. Specifically, we compute the pixel-level BCE loss and select pixels with lower BCE losses for backpropagation in both the dice and BCE loss calculations. Since most generation errors occur at the pixel level, we do not apply sample selection for the mask-wise class prediction in the mask prediction head.

We maintain the same model architecture as Mask2Anomaly, which includes a ResNet-50 [19] backbone, a pixel decoder, a Transformer [47] decoder, and a global mask attention mechanism that independently distributes attention between foreground and background. As in Mask2Anomaly, we keep the ResNet backbone frozen while training the remaining model components, and we employ the

---

<sup>4</sup><https://github.com/l1llyasviel/ControlNet/tree/main>

AdamW [34] optimizer with its default learning rate and scheduler. For our method, the uncertainty loss margins are set to  $\lambda_1 = 0.7$ ,  $\lambda_2 = 0.5$ , and  $\lambda_3 = 0.2$ , and we use a selection ratio  $\alpha = 0.8$ . The loss weights  $\beta_1$  and  $\beta_2$  are set to 10 for both. We use a batch size of 8 for all experiments and train the model on a single NVIDIA A40 48GB GPU.

#### A.4 Training Details on DeepLabv3+ Backbone

For the DeepLabv3+ backbone, we follow the setup in RPL [31], using DeepLabv3+ with WideResNet38 pretrained by Nvidia. The backbone remains fixed, and only the ASPP layers are fine-tuned. We use the Adam [25] optimizer with a learning rate of  $1.0e-6$ . For our method, the contrastive margins  $\lambda_1, \lambda_2, \lambda_3$  are set to 10, 5, 5, the selection ratio is 0.8. The loss weights are 50 and 10 for  $\beta_1$  and  $\beta_2$  respectively. The batch size is set to 8, and all experiments are conducted on two NVIDIA A40 48GB GPUs.

#### A.5 Rule-Based Augmentation

As a typical approach to domain generalization, we implement Rule-Based Augmentation as outlined in [45], using a set of image transformations. Specifically, we apply the following transformations, with their application probabilities indicated in parentheses: color jittering (0.5), Gaussian blur (0.5), random sharpness adjustment (0.5), random contrast adjustment (0.5), random equalization (0.5), random resizing (0.5), random rotation (0.5), random horizontal flipping (0.75), and random cropping (1.0).

### B Additional Ablation Studies

#### B.1 Impact of Hyperparameters

**Loss Margins** Our relative contrastive loss 4 includes three terms, each with a margin value  $\lambda$  controlling the distance penalty limits. These margins are set based on the average uncertainty scores from the training set. Specifically, we compute the differences in uncertainty scores between unknown vs. original known data, unknown vs. augmented known data, original known vs. augmented known data, and set the differences as margins for these distance respectively. Moreover, our two-stage training framework first trains the uncertainty function based on the existing model, allowing this function to adapt to different scales. This provides flexibility in parameter setting even without prior knowledge.

In Table 6, we evaluate the model’s robustness across a wide range of hyperparameter variations. Our default loss margins are  $[\lambda_1, \lambda_2, \lambda_3] = [10, 5, 5]$ . We start by scaling them by 0.1 and 10 and conduct experiments using (1, 0.5, 0.5) and (100, 50, 50), respectively. As shown in Table 6(a), the uncertainty function adjusted to these scales with minimal impact on the results. Further analysis, such as changing the second and third parameters individually, showed that while the relative sizes of the three contrastive losses have some impact, the effect remains minor (see Table 6(b) and Table 6(c)). Ensuring that the parameters are set within an order of magnitude does not affect the results much.

**Table 6: Impact of Loss Margins.** We examine model robustness across various loss margins by evaluating margin scale impacts in (a) and analyzing effects of individual margins in (b) and (c). Results are reported on SMIYC-RA Val(AP & FPR) and MUAD (mIoU) using the DeepLabv3+ architecture.

| (a) Impact of margin scales. |       |               |                  |                 | (b) Impact of margin $\lambda_2$ . |               |                  |                 | (c) Impact of margin $\lambda_3$ . |               |                  |                 |
|------------------------------|-------|---------------|------------------|-----------------|------------------------------------|---------------|------------------|-----------------|------------------------------------|---------------|------------------|-----------------|
| Margins                      | Scale | AP $\uparrow$ | FPR $\downarrow$ | mIoU $\uparrow$ | $\lambda_2$                        | AP $\uparrow$ | FPR $\downarrow$ | mIoU $\uparrow$ | $\lambda_3$                        | AP $\uparrow$ | FPR $\downarrow$ | mIoU $\uparrow$ |
| [1,0.5,0.5]                  | x 0.1 | 93.81         | 5.16             | <b>32.50</b>    | 3                                  | 92.34         | <b>3.27</b>      | <b>31.75</b>    | 3                                  | 92.53         | <b>3.27</b>      | <b>32.24</b>    |
| [10, 5, 5]                   | x 1   | 93.82         | 3.94             | 31.33           | 5                                  | <b>93.82</b>  | 3.94             | 31.33           | 5                                  | <b>93.82</b>  | 3.94             | 31.33           |
| [100,50,50]                  | x 10  | <b>95.73</b>  | <b>2.17</b>      | 29.61           | 7                                  | 91.52         | 4.27             | 31.28           | 7                                  | 92.45         | <b>3.21</b>      | 31.73           |
|                              |       |               |                  |                 | 10                                 | 91.75         | 4.85             | 30.91           | 10                                 | 92.08         | 3.96             | 31.56           |

**Selection Ratio** We determine the selection ratio for our sample selection process by visualizing the selection map of a small batch of data under several choices to ensure that visibly incorrect patterns are removed. To examine the impact of selection ratio to our method, we conducted experiments with selection ratios ranging from 0.6 to 0.9, as detailed in Figure 5 (a). The results show that while including too many pixels (1.0) introduces noise, and including too few (0.6) removes useful regions, the model performance is stable within a wide range (0.7 to 0.9), demonstrating the robustness of the model to this hyperparameter.

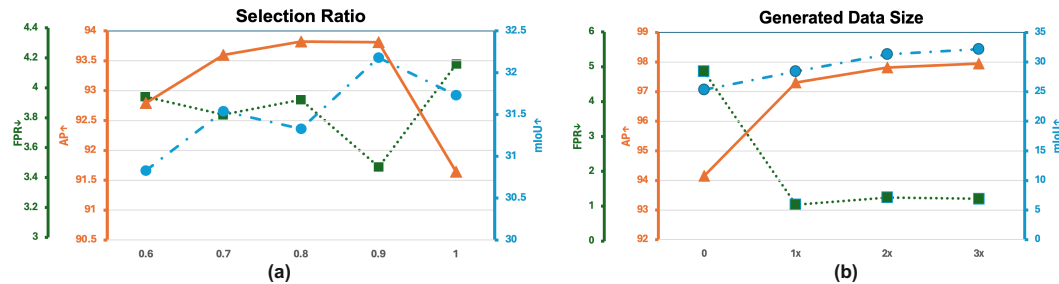


Figure 5: (a) **Impact of Sample Selection Ratio.** We report both anomaly segmentation performance (AP $\uparrow$ , FPR $\downarrow$  on SMIYC-RA Val) and known class segmentation performance (mIoU $\uparrow$  on MUAD). Experiments are conducted under DeepLab v3+ architecture. (b) **Impact of Generated Data Size.** We observe an improvement of performance with the increase of generated data size with the same evaluation under Mask2Former architecture.

## B.2 Impact of the Size of Generated Dataset

By default, we generate our distribution-shift dataset at the same size as Cityscapes. To analyze the impact of the generated dataset size, we scale it to 2x and 3x the size of the Cityscapes training set. As shown in Fig. 5(b), there is a significant improvement from dataset sizes 0 to 1, demonstrating the effectiveness of our generated data, with further gains observed as the dataset size increases.

## B.3 Impact of CG-Aug and Training Strategy for RPL

We evaluate the decoupled contributions of our data augmentation and training strategies with RPL [31] in Table 7. Similar to Table 3, we first replace its original OOD data, which utilizes cut-and-pasted COCO images, with our proposed CG-Aug in Row#2. However, we find the improvement is not as significant. This may be due to certain aspects of RPL’s loss and training design being less suitable for our scenario. Firstly, RPL relies on the original network’s predictions to supervise a learnable residual part. Since the original network does not generalize well to data with domain-shift, this results in imprecise supervision. Secondly, the RPL uncertainty loss focuses solely on increasing uncertainty for unknowns, without adequately addressing the known classes, particularly for augmented images. Additionally, restricting the trainable parameters to a residual block may limit the model’s ability to learn more complex patterns, thereby reducing overall effectiveness.

Next, we replace their training strategy with ours, leading to consistent performance improvement. Those results demonstrate that effectively utilizing the generated training data with multiple distribution shifts remains an open question. Our work takes a step towards analyzing the shortcomings of existing training designs, offering novel and effective strategies for better handling this data.

# C Additional Quantitative Results

## C.1 Additional Results on ACDC-POC

**Performance under Individual Domain Shifts** In addition to the main Table 2, we present the ACDC-POC results with domain-specific splits to provide a more detailed analysis of our method across different types of domain shifts. As shown in Table 8, our method outperforms previous state-of-the-art methods (RPL [31] and Mask2Anomaly [42]) across four domains—fog, rain, snow, and night—on most metrics in both OOD detection and known-class segmentation.

Table 7: **Impact of CG-Aug and Training Strategy.** We evaluate our proposed coherent generative-based augmentation on the previous OOD method, RPL [31], the improvement is not obvious. However, with our training strategy, the performance has largely improved. This demonstrates our training method can effectively utilize the generated training data with multiple distributions.

|          |         | RoadAnomaly  |                     | SMIYC-RA Val |                     | SMIYC-RO Val |                     |
|----------|---------|--------------|---------------------|--------------|---------------------|--------------|---------------------|
| Training | Aug.    | AP↑          | FPR <sub>95</sub> ↓ | AP↑          | FPR <sub>95</sub> ↓ | AP↑          | FPR <sub>95</sub> ↓ |
| RPL [31] | Default | 71.61        | 17.74               | 88.55        | 7.18                | <b>96.91</b> | <b>0.09</b>         |
| RPL [31] | Ours    | 72.46        | 21.85               | 83.50        | 23.88               | 93.30        | 0.51                |
| Ours     | Ours    | <b>74.60</b> | <b>16.08</b>        | <b>93.82</b> | <b>3.94</b>         | 95.20        | 0.19                |

Table 8: **Results on ACDC-POC Datasets with specific domain-shift types:** Fog, Rain, Snow and Night. Our model achieves the best performance in both anomaly segmentation (AP↑, FPR↓) and domain-generalized segmentation (mIoU↑, mAcc↑).

| Method             | Fog         |                     |             |             | Rain        |                     |             |             | Snow        |                     |             |             | Night       |                     |             |             |
|--------------------|-------------|---------------------|-------------|-------------|-------------|---------------------|-------------|-------------|-------------|---------------------|-------------|-------------|-------------|---------------------|-------------|-------------|
|                    | AP↑         | FPR <sub>95</sub> ↓ | mIoU↑       | mAcc↑       | AP↑         | FPR <sub>95</sub> ↓ | mIoU↑       | mAcc↑       | AP↑         | FPR <sub>95</sub> ↓ | mIoU↑       | mAcc↑       | AP↑         | FPR <sub>95</sub> ↓ | mIoU↑       | mAcc↑       |
| RPL [31]           | 88.3        | <b>0.6</b>          | <b>71.2</b> | 93.4        | 67.4        | 1.9                 | 50.9        | 88.0        | 75.1        | 2.5                 | 49.0        | 84.5        | 74.3        | <b>0.8</b>          | 24.4        | 51.2        |
| Ours (DeepLabv3+)  | <b>89.7</b> | 0.7                 | 69.1        | <b>94.3</b> | <b>73.6</b> | <b>1.7</b>          | <b>59.9</b> | <b>92.5</b> | <b>77.9</b> | <b>2.0</b>          | <b>55.4</b> | <b>89.3</b> | <b>83.7</b> | <b>0.8</b>          | <b>35.4</b> | <b>65.3</b> |
| M2A [42]           | 83.9        | <b>0.9</b>          | 67.3        | <b>94.3</b> | 75.9        | 1.7                 | 53.2        | 91.0        | 71.0        | 2.3                 | 45.2        | 86.4        | 75.8        | 3.9                 | 29.5        | 61.8        |
| Ours (Mask2Former) | <b>90.5</b> | <b>0.9</b>          | <b>69.7</b> | 94.2        | <b>91.3</b> | <b>0.3</b>          | <b>54.5</b> | <b>91.2</b> | <b>90.7</b> | <b>0.6</b>          | <b>51.5</b> | 86.7        | <b>88.7</b> | <b>0.4</b>          | <b>31.8</b> | 61.6        |

**Per-Class Segmentation Results** We evaluated per-class segmentation results and compared them with the baseline DeepLabv3+ [7] model. Results are presented in Table 9. Our method improves segmentation performance (mIoU) across most categories. However, performance in fence, pole, and traffic sign remains similar (with differences of less than 1%), and performance on vegetation decreases by 3%, likely due to poor generation quality for this class.

Table 9: **Per-class segmentation results.** We present the segmentation performance (mIoU) for each known class on the ACDC-POC dataset. Compared to the baseline model (DeepLabv3+ [7]), our method improves performance in most categories.

| Method         | road         | sidewalk     | building     | wall         | fence        | pole         | traffic light | traffic sign |
|----------------|--------------|--------------|--------------|--------------|--------------|--------------|---------------|--------------|
| DeepLabv3+ [7] | 77.57        | 37.7         | 63.55        | 17.46        | <b>31.22</b> | 49.42        | 64.24         | 52.78        |
| Ours           | <b>85.75</b> | <b>58.16</b> | <b>74.8</b>  | <b>40.79</b> | 29.43        | <b>50.81</b> | <b>71.69</b>  | <b>53.01</b> |
|                |              |              |              |              |              |              |               |              |
|                | vegetation   | terrain      | sky          | person       | rider        | car          | truck         | bus          |
| DeepLabv3+ [7] | <b>75.75</b> | 10.58        | 81.12        | 54.17        | 19.37        | 77.89        | <b>50.76</b>  | <b>50.74</b> |
| Ours           | 72.82        | <b>30.21</b> | <b>81.35</b> | <b>62.77</b> | <b>32.17</b> | <b>79.99</b> | 49.41         | 49.33        |
|                |              |              |              |              |              |              |               |              |
|                | train        | motorcycle   | bicycle      |              |              |              |               |              |
| DeepLabv3+ [7] | 30.33        | 13.77        | 22.05        |              |              |              |               |              |
| Ours           | <b>38.08</b> | <b>29.19</b> | <b>38.52</b> |              |              |              |               |              |

## C.2 Comprehensive Metric Results on SMIYC

To provide a comprehensive analysis of our method, we complement the main metrics (AP, FPR) with component-wise metrics (sIoU, PPV, F1) on the SMIYC benchmark. The results are summarized in Table 10. As shown, our method outperforms RPL [31] and Mask2Anomaly [42] across most evaluation metrics.

## C.3 Comparison with POC [12]

We present additional comparison results of our CG-Aug method against POC [12] across six datasets. Following the experimental setup in POC [12], we replace the default OOD data (COCO) in Mask2Anomaly [42] with our CG-Aug. As shown in Table 11, our method outperforms both POC variations on most datasets. Notably, for FS-Static [4], where OOD objects are introduced through



Table 10: **Comprehensive Metric Results on SMIYC.** We present the results of our method across pixel-wise metrics (AP, FPR) and component-wise metrics (sIoU, PPV, and F1). Compared to recent methods RPL [31] and Mask2Anomaly [42], our method achieves superior or comparable results.

| Method   | Backbone    | SMIYC-RA     |                     |              |              |              | SMIYC-RO     |                     |              |              |              |
|----------|-------------|--------------|---------------------|--------------|--------------|--------------|--------------|---------------------|--------------|--------------|--------------|
|          |             | AP↑          | FPR <sub>95</sub> ↓ | sIoU↑        | PPV↑         | F1↑          | AP↑          | FPR <sub>95</sub> ↓ | sIoU↑        | PPV↑         | F1↑          |
| RPL [31] | DeepLab v3+ | 83.49        | 11.68               | 49.76        | 29.96        | 30.16        | 85.93        | 0.58                | <b>52.61</b> | 56.65        | 56.69        |
| Ours     |             | <b>88.06</b> | <b>8.21</b>         | <b>56.15</b> | <b>34.66</b> | <b>37.83</b> | <b>90.71</b> | <b>0.26</b>         | 48.13        | <b>66.73</b> | <b>58.02</b> |
| M2A [42] | Mask2Former | 88.70        | 14.60               | <b>60.40</b> | 45.70        | 48.60        | 93.30        | 0.20                | <b>61.40</b> | 70.30        | <b>69.80</b> |
| Ours     |             | <b>91.92</b> | <b>7.94</b>         | 58.74        | <b>45.77</b> | <b>48.74</b> | <b>95.29</b> | <b>0.07</b>         | 59.43        | <b>73.51</b> | 68.70        |

cut-and-paste, COCO achieves the best performance; however, this setup lacks the ability to reflect realistic OOD objects encountered in real-world scenarios.

Table 11: **Comparison of our CG-Aug with POC.** Following the experimental setup in POC [12], we replace the default OOD data (COCO) in Mask2Anomaly [42] with our CG-Aug, and compare with two versions of POC. Our method outperforms both POC variations on most datasets.

|               | Road Anomaly |              | SMIYC-RA21 (val) |             | SMIYC-RO21 (val) |            | ACDC-POC     |             | FS-L&F (val) |            | FS-Static (val) |             |
|---------------|--------------|--------------|------------------|-------------|------------------|------------|--------------|-------------|--------------|------------|-----------------|-------------|
| Mask2Anomaly+ | AP↑          | FPR↓         | AP↑              | FPR↓        | AP↑              | FPR↓       | AP↑          | FPR↓        | AP↑          | FPR↓       | AP↑             | FPR↓        |
| COCO          | 79.70        | <b>13.45</b> | 94.50            | 3.30        | 88.6             | 0.30       | 73.77        | 3.60        | 69.41        | 9.46       | <b>90.54</b>    | <b>1.98</b> |
| POC-c [12]    | 82.3         | 36.7         | 93.8             | 2.1         | 95.3             | 0.3        | 74.5         | 7.6         | 68.8         | 11.4       | 87.4            | 3.1         |
| POC-alt [12]  | 78.0         | 24.6         | 92.1             | 8.4         | <b>96.0</b>      | <b>0.1</b> | 72.0         | 8.4         | 73.0         | <b>9.2</b> | 87.0            | 2.1         |
| CG-Aug (Ours) | <b>85.47</b> | 22.38        | <b>97.96</b>     | <b>1.55</b> | 89.80            | 0.12       | <b>86.17</b> | <b>1.05</b> | <b>76.56</b> | 10.17      | 85.70           | 7.16        |

#### C.4 Comparison with DG Methods on the Original ACDC Dataset

To assess the effectiveness of our method in domain generalization, we conducted additional comparisons with several recent approaches, including IBN [39], IterNorm [22], IW [40], ISW [10], ISSA [27], and CMFormer [3], on the ACDC [44] dataset. As shown in Table 12, our method outperforms all ResNet-based methods in the Fog, Rain, and Snow domains, achieving comparable results in the Night domain. Among Mask2Former-based methods, our approach also surpasses ISSA [27], which similarly uses a ResNet backbone. However, there remains a notable performance gap between our method and CMFormer [3], likely due to CMFormer’s use of the Swin Transformer [32] as the backbone for Mask2Former.

Table 12: **Domain generalization performance comparison between our method and other DG methods.** Results from other methods are taken from CMFormer [3]. All methods are trained on the Cityscapes [11] dataset and tested on the ACDC [44] dataset. Results are shown in mIoU (%).

| Method       | Backbone    | Fog          | Night       | Rain         | Snow         | Mean         |
|--------------|-------------|--------------|-------------|--------------|--------------|--------------|
| IBN[39]      | ResNet      | 63.8         | 21.2        | 50.4         | 49.6         | 43.7         |
| Iternorm[22] |             | 63.3         | 23.8        | 50.1         | 49.9         | 45.3         |
| IW[40]       |             | 62.4         | 21.8        | 52.4         | 47.6         | 46.6         |
| ISW[10]      |             | 67.5         | <b>33.2</b> | 55.9         | 53.2         | 52.5         |
| Ours         |             | <b>74.84</b> | 30.33       | <b>61.15</b> | <b>57.80</b> | <b>55.95</b> |
| ISSA[27]     | Mask2Former | 67.5         | 33.2        | 55.9         | 53.2         | 52.5         |
| CMFormer[3]  |             | <b>77.8</b>  | <b>33.7</b> | <b>67.6</b>  | <b>64.3</b>  | <b>60.1</b>  |
| Ours         |             | 70.85        | 29.39       | 54.54        | 52.97        | 51.61        |

## D Visualization of Generated Data

In Fig. 6, we provide additional visualization examples of our generated images (row 2), corresponding selection maps (row 3), and the loss maps used to produce the selection map (row 4). Below each column, we display the weather, time, and location prompts that guide the model in generating

diverse covariate shifts, as well as the OOD prompts used to generate objects. As shown, our method effectively generates images with both domain and semantic shifts, with the novel objects blending seamlessly into the background (e.g. pose and lighting). Additionally, our sample selection process effectively filters out some generation errors (highlighted in red boxes).



Figure 6: **Visualization of Generated Images.** Row 1: Original images from Cityscapes. Row 2: Generated images featuring both semantic and domain shifts. Row 3: Selection map used to calculate selected cross-entropy loss during training. Row 4: Cross-entropy loss map used to produce the selection map (excluding the OOD regions, which are not involved in known class segmentation loss calculation). Below each column, we display the weather, time, and location prompts that guide the model in generating diverse covariate shifts, along with the OOD prompts for object generation. Red boxes highlight generation errors.

## E Discussion of Generation Failures and Their Impact

A limitation of our method is its reliance on the quality of the generative model. Although we apply offline auto-filtering and online sample selection to minimize the impact of generation failures during training, some issues may still arise. Specifically, we observe that generation failures typically occur in the following scenarios: (a) remote scenes, (b) small objects, and (c) text-related elements. These limitations highlight the current constraints of generative models and suggest areas for future research. Below, we discuss the impact of generation failures:

**Impact on Class-Specific Learning:** Generation failures can adversely affect specific classes. As shown in Table 9, we evaluated per-class segmentation results and compared them with the baseline model on Cityscapes. Performance in categories—such as fence, pole, and traffic sign—remains similar (differences of less than 1%), and vegetation shows a 3% decrease, likely due to lower generation quality for these classes.

**Performance Saturation:** We observe that performance tends to saturate with increasing amounts of generated data. Experiments with dataset scaling from 1.0x to 2.0x and 3.0x Cityscapes sizes, as shown in Figure 5, indicate that while performance improves with larger dataset size, it eventually plateaus. This saturation may result from an interplay between the benefits of additional data and the adverse effects of generation failures.

## F Societal Impacts

Enhancing OOD detection in autonomous vehicles can significantly improve safety by enabling these systems to better recognize and respond to novel and unexpected situations, thereby reducing the risk of accidents. Improved robustness to domain shifts also contributes to greater resilience and safety across diverse driving scenarios. However, improved OOD detection may lead to an over-reliance on autonomous systems, potentially reducing the vigilance of human drivers or passengers in semi-autonomous vehicles. Additionally, unintended biases in OOD detection systems could result in unsafe responses to certain situations, particularly if training data does not sufficiently cover diverse scenarios, potentially compromising safety in rare but critical cases.

## 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

## 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

## 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

#### 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

#### 13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:



Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

#### 14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

#### 15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.