CSL: Class-Agnostic Structure-Constrained Learning for Segmentation Including the Unseen

Hao Zhang[1], **Fang Li**[1], **Lu Qi**[2], **Ming-Hsuan Yang**[2, 3], **Narendra Ahuja**[1]
[1]University of Illinois at Urbana-Champaign
[2]University of California Merced
[3]Google Research
{haoz19, fangli3}@illinois.edu, {lqi5, mhyang}@ucmerced.edu, n-ahuja@illinois.edu

Abstract

Addressing Out-Of-Distribution (OOD) Segmentation and Zero-Shot Semantic Segmentation (ZS3) is challenging, necessitating segmenting unseen classes. Existing strategies adapt the class-agnostic Mask2Former (CA-M2F) tailored to specific tasks. However, these methods cater to singular tasks, demand training from scratch, and we demonstrate certain deficiencies in CA-M2F, which affect performance. We propose the Class-Agnostic Structure-Constrained Learning (CSL), a plug-in framework that can integrate with existing methods, thereby embedding structural constraints and achieving performance gain, including the unseen, specifically OOD, ZS3, and domain adaptation (DA) tasks. There are two schemes for CSL to integrate with existing methods (1) by distilling knowledge from a base teacher network, enforcing constraints across training and inference phrases, or (2) by leveraging established models to obtain per-pixel distributions without retraining, appending constraints during the inference phase. We propose soft assignment and mask split methodologies that enhance OOD object segmentation. Empirical evaluations demonstrate CSL's prowess in boosting the performance of existing algorithms spanning OOD segmentation, ZS3, and DA segmentation, consistently transcending the state-of-art across all three tasks.

Introduction

Semantic segmentation is a fundamental task in computer vision, which associates with each pixel in a given image probabilities of belonging to different classes. Recent approaches have achieved remarkable results on several closed-set benchmarks that contain images from known classes, called In-Distribution (ID) images. However, the segmentation with the unseen, e.g., Out-Of-Distribution (OOD), Zero-shot-semantic (ZS3) segmentation, is always challenging because it requires segmentation and discrimination based on training on only ID images.

The existing methods for such tasks can be distinguished by whether they use OOD data for training. Some methods expand the training set to include OOD images from other datasets (Hendrycks, Mazeika, and Dietterich 2018a; Chan, Rottmann, and Gottschalk 2021a; Kang, Kwak, and Kang 2022; Tian et al. 2022), or utilize large-scale models, e.g., SAM (Kirillov et al. 2023) to generate region proposals. Such expansion-based approaches are not of great interest in this paper since we aim to solve the general problem of OOD segmentation without having access to any OOD images for training. We propose to learn models of objects that extend to classes beyond those in the ID set.

Existing segmentation methods typically infer OOD if some properties of the outputs are sufficiently different from those seen on ID images. An example of properties used is the uncertainty in pixel label prediction, as in the SML methods (Figure 2). Other examples of properties used are errors in image reconstruction (Lis et al. 2019), and the similarity to which results are perturbed by adversarial attacks (Liang, Li, and Srikant 2017; Besnier et al. 2021). However, these methods result in noisy predictions due to a lack of structured knowledge. Current techniques, such as those in (Nayal et al. 2023; Grcić, Šarić, and Šegvić 2023), tackle this issue using the region-based framework, Mask2Former (M2F) (Cheng et al. 2021). However, to achieve optimal performance, they necessitate OOD data and complete train the model from scratch. For ZS3 or open-word semantic segmentation, existing methods (Ding et al. 2022; Xu et al. 2022) typically leverage CA-M2F, trained on the ID set, as a region generator and utilize CLIP (Radford et al. 2021) to identify the semantic class for each region. Some works (Qi et al. 2022) empirically demonstrate that CA training benefits the performance on OOD data and since M2F decouples the per-pixel prediction task into 2 sub-tasks: (1) mask prediction and (2) per-mask class prediction optimized by the mask loss and class loss, a straightforward way is removing the class loss and leverage hard assignment as post-processing during inference. However, our observations reveal that such adjustments are insufficient in eliminating class information. The process of hard assignment frequently results in unanticipated outcomes. For instance, certain objects might not have corresponding masks, and in certain situations, multiple objects may be erroneously blended into a single mask.

In this paper, we present Class-Agnostic Structure-Constrained Learning (CSL) framework for seamless integration with existing methodologies, including OOD, ZS3, and DA segmentation, to improve their performance by incorporating structure constraints. CSL offers two plug-in integration schemes: (1) Knowledge distillation from a base

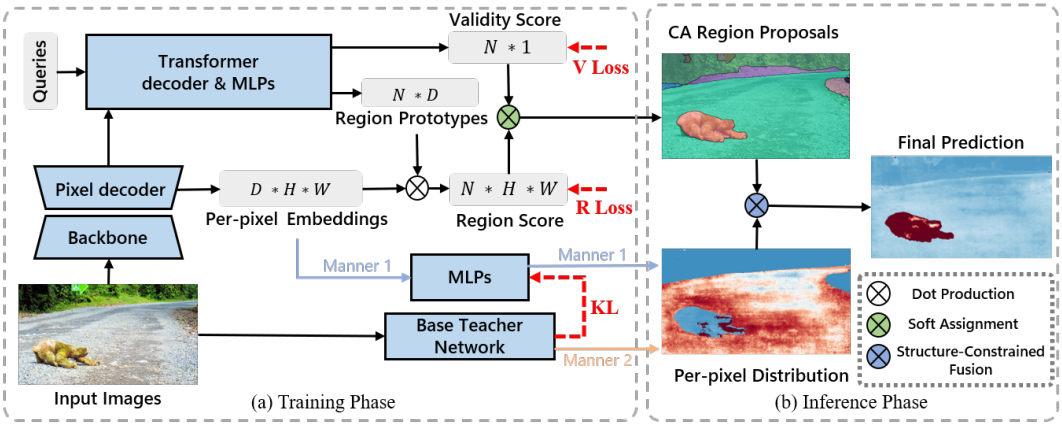

Figure 1: Overview of our CSL framework. CSL consists of a backbone, a pixel decoder, a transformer decoder, a base teacher network, and MLPs. N learnable region queries and the image features are fed to the transformer decoder and MLPs, to obtain N pairs of latent region prototypes and their validity scores. We calculate the normalized similarity between each element $\mathcal{E}_{h,w}$ of per-pixel embeddings and each $\mathbf{P}_n, n \in \{1, 2, ...N\}$ by simple dot production followed by the sigmoid function to get N region scores. The validity scores indicate the degree of the region prototypes are valid for given images. During training, a valid loss, a region loss, and a distillation loss are used to optimize the model. Instead of assigning each pixel from the input image to one of the prior fixed classes, CSL assigns it to one of N learnable region prototypes by our proposed soft assignment. During inference, we introduce structure-constrained Fusion to calculate the final prediction.

teacher network, potentially any existing method, with structure constraints imposed during training and inference. (2) Direct application of existing methods for per-pixel prediction, incorporating structure constraints solely during inference, bypassing retraining. While the first style facilitates end-to-end training, the second negates the need for retraining, and both surprisingly yield comparable gains over foundational methods. In semantic segmentation, annotations commonly amalgamate all instances of a class into a singular mask. We split this mask into multiple isolated components for training, mitigating bias from seen classes. During inference, CSL employs a soft assignment to derive region proposals at the disconnected-component level. Compared with the prevalent hard assignment, the soft assignment boosts the performance of unseen samples. The main contributions of this paper are as follows:

- We present CSL, a modular plug-in framework with 2 2 schemes, designed for seamless integration with established methodologies, enhancing the segmentation of unseen classes by incorporating structural constraints.
- We propose mask split preprocessing, splitting class masks into isolated components, effectively attenuating the bias of seen class data. Furthermore, we employ a soft assignment in post-inference for region proposal generation and elucidate the driving factors behind the observed performance enhancements.
- Through extensive experimental validation, we ascertain that CSL markedly enhances 10 prevailing techniques across all three segmentation tasks, including OOD segmentation, ZS3, and DA segmentation, consistently outstripping state-of-the-art benchmarks.

Related Work

Out-of-Distribution Segmentation

Uncertainty-based Methods. Leveraging pixel-wise prediction uncertainty, OOD segmentation methods (Hendrycks and Gimpel 2016; Lee et al. 2017; Liang, Li, and Srikant 2017; Tian et al. 2021; Zhang and Zhang 2022) avoid retraining, thus saving computation. However, issues arise in hard-predicted regions. Jung et al. (Jung et al. 2021) refine boundary anomaly scores, while others (Kendall and Gal 2017; Lakshminarayanan, Pritzel, and Blundell 2017; Mukhoti and Gal 2018) apply MC dropout, often with limited success. **Image Reconstruction**. Autoencoders and GANs dominate reconstruction methods (Baur et al. 2019; Creusot and Munawar 2015; Di Biase et al. 2021; Haldimann et al. 2019; Liu et al. 2020). Notably, ID-only trained models (Xia et al. 2020; Lis et al. 2020; Ohgushi, Horiguchi, and Yamanaka 2020; Creusot and Munawar 2015; Lis et al. 2019; Vojir et al. 2021) effectively reconstruct ID samples, but falter with OODs, hindered further by domain sensitivity and extended training/inference times. **Adversarial Attacks**. Adversarial attacks serve as OOD data simulators in image classification (Goodfellow, Shlens, and Szegedy 2014) and detection (Ma et al. 2018). Besnier et al.'s ObsNet, though utilizing Local Adversarial Attacks, faces noisy prediction challenges due to structural information absence. **Outlier Exposure**. The outlier exposure (OE) strategy by Hendrycks et al. (Hendrycks, Mazeika, and Dietterich 2018b) augments the training set with non-overlapping outliers. Conversely, some methods (Chan, Rottmann, and Gottschalk 2021b; Bevandić et al. 2019; Vandenhende et al. 2020; Bevandić et al. 2018; Liu et al. 2022; Tian et al. 2022) embed OOD objects from datasets like COCO (Lin et al. 2014) and ADE20K (Zhou

et al. 2019), potentially reducing OOD segmentation to mere binary segmentation due to overlaps.

Proposed Method

As shown in Figure 1, CSL provides two schemes to plug in existing methods. The first is an end-to-end scheme, which distills the knowledge from the base teacher network to the CSL framework. The second scheme directly utilizes the existing models as a base teacher network to obtain per-pixel distributions and fuse them with the class-agnostic region proposals during inference without retraining them. Validity loss, region loss, and distill loss, which is the mean Huber loss between the predicted per-pixel distribution and the output of the base teacher network (only for scheme$_1$), are used for optimization.

Class-Agnostic Training. To capture the essential features of semantic classes that are applicable beyond the training classes, we design CSL in a class-agnostic way. It learns region prototypes characterized by visual appearance and spatial features and uses them to generate region proposals. CSL firstly uses a backbone and a pixel decoder to generate multi-level feature embeddings $\mathcal{E}^l \in \mathbb{R}^{H_l \times W_l \times D_l}$, where $l \in \{4, 8, 16, 32\}$ indicating the downsampled size of feature map compared to the original image. D_l is the dimension of the embeddings. In addition, we have N learnable queries, which cascadely interact with multi-level feature embeddings $\mathcal{E}^l$, where $l \in \{8, 16, 32\}$, to generate N region prototypes $\mathbf{P} \in \mathbb{R}^{N \times 256}$. These prototypes act as centers for grouping the per-pixel embeddings $\mathcal{E} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 256}$ and follows by a upsampling to get the region prediction $\mathbf{R} \in [0, 1]^{N \times H \times W}$. The region scores $\mathbf{R}$, and the validity scores $\mathbf{V} \in [0, 1]^N$ are fed into the soft assignment module (Sec 3.3) and generate region proposals.

Comparison with CA-M2F. In Mask2Former (Cheng et al. 2021), they use the semantic class predictions $\mathbf{V} \in [0, 1]^{N \times C}$ instead of the validity score, where C is the number of classes, and it empirically yields exceptional semantic segmentation results $\mathbf{Y} \in \mathbb{R}^{H \times W \times C}$ by matrix multiplication between $\mathbf{V}$ and $\mathbf{R}$. We explain this matrix multiplication as the calculation of the likelihood $p(x_{h,w} \in c)$, where $x_{h,w}$ and c denote the pixel at location (h, w) of input image $\mathbf{X}$ and the class $c \in C$:

$$\begin{aligned} \max_c p(x_{h,w} \in c) &= \max_c \sum_{n=1}^{N} p(\, x_{h,w} \in \mathcal{R}_n \cap x_{h,w} \in c) \\ &= \max_c \sum_{n=1}^{N} p(x_{h,w} \in \mathcal{R}_n) \times p(x_{h,w} \in c | x_{h,w} \in \mathcal{R}_n) \\ &= \max_c \sum_{n=1}^{N} r_{n,h,w} \times v_{n,c} = \max_c (\mathbf{R}^\top \cdot \mathbf{V})_{h,w,c}, \end{aligned} \tag{1}$$

where the $r_{n,h,w}$ and $v_{n,c}$ at $\mathbf{R}$ and $\mathbf{V}$ indicate the probability of pixel $x_{h,w}$ belonging to region $\mathcal{R}_n$ and region $\mathcal{R}_n$ belonging to class c. Given that the pixels in the same region follow the same class distribution, we have $p(\mathcal{R}_n \in c) = p(x_{h,w} \in c | x_{h,w} \in \mathcal{R}_n)$.

CA-M2F removes the class loss and uses $\mathbf{R}$ as the region proposals, which are demonstrated to be unsatisfactory due to the redundancy of regions. To reduce useless regions, some methods keep the class prediction but reduce it to a binary classification indicating the validity of the region. And they utilize hard assignments during inference to generate the region proposals.

Soft Assignment. However, the hard assignment requires a manually selected threshold, which limits the generalization, and multiple experiments demonstrate its limited performance on OOD objects (Figure 3). Thus, we creatively propose a soft assignment module to maximize the objective of the class-agnostic semantic segmentation: $p(x_{h,w} \in \mathcal{R}_n \cap \mathcal{R}_n \in \mathcal{V})$, where $\mathcal{R}_n \in \mathcal{V}$ denotes region $\mathcal{R}_n$ being valid. In CSL, we use binary cross-entropy losses for optimal validity score and region score by maximizing the likelihoods of $p(x_{h,w} \in \mathcal{R}_n)$ and $p(x_{h,w} \in \mathcal{V} | x_{h,w} \in \mathcal{R}_n)$. This allows us to interpret v_n, in the validity score matrix $\mathbf{V}$, as the probability of the pixels in the region $\mathcal{R}_n$ being valid, which we call as the region's validity score. Since multiple overlapping region prototypes exist, $\mathbf{V}$ helps select the valid masks. For instance, simple images with few objects result in fewer valid regions than complex ones. Similarly, we can interpret $r_{n,h,w}$, in the region prediction matrix $\mathbf{R}$, as the probability that pixel $x_{h,w}$ belongs to region $\mathcal{R}_n$, which we call as the region score. The objective for class-agnostic segmentation can be derived from these two likelihoods as follows:

$$\begin{aligned} &\max_n p(x_{h,w} \in \mathcal{R}_n \cap \mathcal{R}_n \in \mathcal{V}) \\ &= \max_n p(x_{h,w} \in \mathcal{R}_n) \times p(x_{h,w} \in \mathcal{V} \mid x_{h,w} \in \mathcal{R}_n) \\ &= \max_n (r_{n,h,w} \times v_n) = \max_n (\mathbf{R}^\top * \mathbf{V})_{h,w,n}, \end{aligned} \tag{2}$$

which maximizes the probability that the pixel $x_{h,w}$ is from region $\mathcal{R}_n$ and the region $\mathcal{R}_n$ is valid, where * denotes pixel-wise multiplication. Note that given that pixel $x_{h,w}$ belongs to region $\mathcal{R}_n$, the validity of $\mathcal{R}_n$ can be represented by $x_{h,w}$ because the validity score is assigned per region.

Comparison with hard assignment. For panoptic segmentation, Mask2Former and EntitySeg employ hard assignment during inference Specifically, a binary region mask is generated from each region prediction by checking at each pixel if the mask score exceeds a certain threshold. The region masks are then stacked in ascending order of validity score, where a mask with a higher validity score covers those with lower scores. However, this approach has limitations. First, hard assignment employs a fixed threshold to filter out low-score regions, which often results in missed pixels that are not allocated to any region (a, c, and d in Figure 3). Second, it performs poorly in complex and detailed scenes as the final results are obtained based on hard region masks instead of per-pixel scores in the soft assignment (f in Figure 3).

Mask Split. Another problem is that existing methods such as entity segmentation (Qi et al. 2022) work well when training with instance-wise labels while failing with semantic-wise labels. We believe and demonstrate it's because the semantic-wise labels still contain class information, which introduces the bias of seen class. Therefore, we

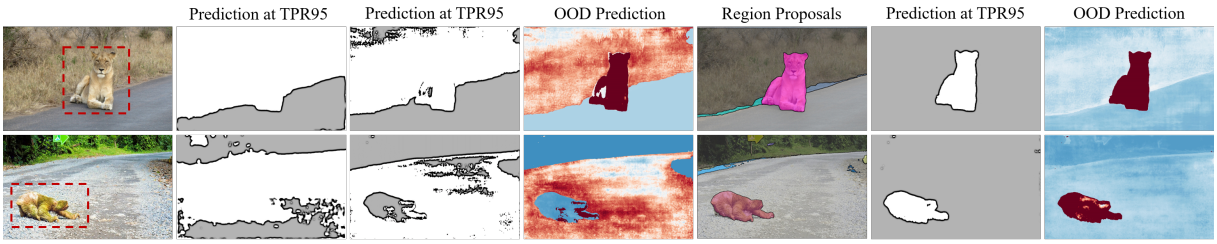

Figure 2: Visualisations for hard mask predictions at TPR=95% and per-pixel Out-of-distribution (OOD) scores. We compare the results of SML, and ObsNet with our proposed CSL. In the hard mask predictions, white and gray indicate being predicted to be OOD. In the OOD scores, the red and blue intensity values correspond to the magnitudes of the OOD scores above and below the decision boundary, respectively. (d) shows the region proposals from our CSL with scheme$_2$.

Method	OOD	Pixel Level		Component Level	
		AUPR	FPR$_{95}$	sIoU gt	mean F1
PEBAL	✓	49.1	40.8	38.9	14.5
ME	✓	85.5	15.0	49.2	28.7
DH	✓	78.0	9.8	54.2	31.1
SynBoost	✓	56.4	61.9	34.7	10.0
IR	×	52.3	25.9	39.7	12.5
ObsNet	×	75.4	26.7	44.2	45.1
+CSL$_1$	×	79.9	**7.1**	46.1	50.2
+CSL$_2$	×	**80.1**	7.2	**46.5**	**50.4**

Table 1: Results on SMIYC-AT.

Method	OOD	Pixel Level		Component Level	
		AUPR	FPR$_{95}$	sIoU gt	mean F1
ME	✓	85.1	0.75	47.9	50.4
DH	✓	80.8	6.02	**48.5**	55.6
SynBoost	✓	71.3	3.15	44.3	37.6
RI	×	54.14	47.1	**57.6**	36.0
IR	×	37.7	4.7	16.6	8.4
DaCUP	×	81.5	1.13	37.7	46.0
+CSL$_1$	×	86.8	0.9	44.3	50.7
+CSL$_2$	×	**87.1**	**0.7**	44.7	**51.0**

Table 2: Results on SMIYC-OT.

present a simple yet effective preprocessing method named mask split. Annotation of semantic segmentation often consists of multiple disconnected regions of the same class in a single mask, which forces the model to predict all pixels from the same class to the same mask, thereby enforcing the embeddings of the pixels from the same semantic class to be similar, which causes the bias. To overcome this limitation, we propose a simple yet effective method called Mask Split to overcome this limitation. We split these two disconnected components as depicted in Supplementary Material. This approach reduces the class information and allows the model to predict instances without being biased toward any particular class.

Structure-Constrained Fusion. Intuitively, making predictions at each pixel independently does not benefit from the predictions at other, nearby pixels, which are correlated.

Method	OOD	Pixel Level		Component Level	
		AUPR	FPR$_{95}$	sIoU gt	mean F1
ME	✓	77.90	9.70	45.90	49.92
SynBoost	✓	81.71	4.64	36.83	48.72
RI	×	82.93	35.75	49.21	52.25
DaCUP	×	81.37	7.36	38.34	51.24
+CSL$_1$	×	83.07	6.88	40.43	51.57
+CSL$_2$	×	83.41	6.92	40.89	51.36
NFlowJS	×	89.28	0.65	54.63	61.75
+CSL$_1$	×	89.48	**0.48**	54.78	62.25
+CSL$_2$	×	**89.79**	0.51	**55.01**	**62.37**

Table 3: Results on LAF NoKnown.

To address this, we introduce structure-constrained rectification (SCF), which utilizes structure constraints, that interrelate predictions at different pixels, to optimize per-pixel predictions. This helps improve performance on multiple tasks. Our proposed approach leverages soft assignment to generate region proposals $\mathcal{R} \in \{0,1\}^{H\times W\times N}$. These proposals, along with the per-pixel distribution $\mathbf{D} \in \mathbb{R}^{H\times W\times C}$ from scheme 1 or scheme 2, are fed into our proposed SCF. For OOD segmentation, we set C to 1 since the prediction is binary, and we only need to consider the probability of belonging to OOD. And, for the domain adaptation (DA) and zero-shot semantic segmentation (ZS3) tasks, C is equal to the number of classes.

Each $\mathbf{D}_c \in \mathbb{R}^{H\times W}$ indicates per-pixel distribution for class c, where $c \in \{1,...,C\}$. We compute the region-wise score as the average of the pixel-wise scores within each region proposal $\mathcal{R}_n \in \{0,1\}^{H\times W}$, where $n \in \{1,...,N\}$. Then we combine the region-wise scores and the pixel-wise scores to obtain the hybrid score $\mathbf{H}$ using the equation:

$$\mathbf{H}_{c,n} = \frac{\sum_{h,w} \mathbf{D}_{c,n} * \mathcal{R}_n}{\sum_{h,w} \mathcal{R}_n} \times \mathbf{D}_{c,n}, \quad n \in \{1,...,N\} \quad (3)$$

where $\mathbf{H}_{c,n}$, $\mathbf{D}_{c,n}$ indicates the hybrid score and per-pixel distribution for class c within region n, $*$ is the pixel-wise multiplication and $\triangle$ indicates pixel-wise multiplication.

Method	OOD Data	SMIYC (AT)-val			SMIYC (AT)-test			Road Anomaly		
		FPR_{95}	AP	AUROC	FPR_{95}	AP	AUROC	FPR_{95}	AP	AUROC
SML†	×	51.0	47.7	81.7	43.33	44.68	86.57	49.63	25.71	81.90
+CSL_2	×	$22.0_{\uparrow 29}$	$55.4_{\uparrow 7.8}$	$88.4_{\uparrow 6.7}$	$39.7_{\uparrow 3.7}$	$47.2_{\uparrow 2.6}$	$87.5_{\uparrow 0.86}$	$\mathbf{41.03}_{\uparrow 8.60}$	$31.78_{\uparrow 6.07}$	$84.77_{\uparrow 2.87}$
IR†	×	-	-	-	32.17	49.36	87.03	69.79	33.43	79.92
+CSL_2	×	-	-	-	$21.7_{\uparrow 10.4}$	$54.5_{\uparrow 5.1}$	$89.3_{\uparrow 2.3}$	$59.16_{\uparrow 10.63}$	$35.48_{\uparrow 2.05}$	$82.45_{\uparrow 2.53}$
ObsNet†	×	40.2	72.7	91.9	61.73	56.91	86.22	64.25	48.13	83.18
+CSL_2	×	$25.0_{\uparrow 15.2}$	$75.5_{\uparrow 2.8}$	$95.2_{\uparrow 3.3}$	$32.1_{\uparrow 29.6}$	$64.8_{\uparrow 7.8}$	$92.24_{\uparrow 6.02}$	$47.21_{\uparrow 17.04}$	$53.24_{\uparrow 5.11}$	$86.92_{\uparrow 3.74}$
ObsNet v2†	×	30.3	74.5	93.7	26.69	75.44	93.80	55.75	54.64	86.78
+CSL_2	×	$\mathbf{5.8}_{\uparrow 24.5}$	$\mathbf{83.6}_{\uparrow 9.1}$	$\mathbf{97.4}_{\uparrow 3.7}$	$\mathbf{7.16}_{\uparrow 19.5}$	$\mathbf{80.1}_{\uparrow 4.6}$	$\mathbf{96.46}_{\uparrow 2.66}$	$43.80_{\uparrow 11.95}$	$\mathbf{61.38}_{\uparrow 6.74}$	$\mathbf{91.08}_{\uparrow 4.3}$

Table 4: Quantitative results on SMIYC (Anomaly Track) and Road Anomaly. We show the results obtained by combining CSL_2 with 3 well-established OOD segmentation methods (indicated by †). The best results are highlighted in bold.

Method	ST	RT	COCO-stuff			PASCAL VOC 2012		
			mIoU(S)	mIoU(U)	hIoU	mIoU(S)	mIoU(U)	hIoU
ZegFormer	×	×	36.6	33.2	34.8	86.4	63.6	73.3
+CSL_2	×	×	$37.5_{\uparrow 0.9}$	$36.2_{\uparrow 3}$	$36.9_{\uparrow 2.1}$	$87.1_{\uparrow 0.7}$	$68.6_{\uparrow 5}$	$76.9_{\uparrow 3.6}$
ZSSeg	×	×	39.3	36.3	37.8	83.5	72.5	77.5
+CSL_2	×	×	$40.1_{\uparrow 0.8}$	$38.3_{\uparrow 2}$	$39.2_{\uparrow 1.4}$	$84.7_{\uparrow 1.2}$	$76.9_{\uparrow 4.4}$	$80.6_{\uparrow 3.1}$
ZegCLIP	×	×	40.2	41.4	40.8	91.9	77.8	84.3
+CSL_2^*	×	×	$40.4_{\uparrow 0.2}$	$42.8_{\uparrow 1.4}$	$41.6_{\uparrow 0.8}$	$92.3_{\uparrow 0.4}$	$79.4_{\uparrow 1.6}$	$85.5_{\uparrow 1.2}$

Table 5: Quantitative results for ZS3 on COCO-stuff and PASCAL VOC benchmarks. The "mIoU(S)", "mIoU(U)", and "hIoU" denote the mIoU of seen classes, unseen classes, and their harmonic mean. "ST" and "RT" denote self-training and re-training.

Experimental Results

Experimental Setup

In all our experiments [1], we utilize ResNet50 as the backbone and FPN as the pixel decoder. All experiments for OOD segmentation are performed without any OOD data. In the DA and ZS3 experiments, we use the same training data as the comparative methods. CSL_1 and CSL_2 represent scheme 1 and 2, and CSL_2 doesn't require retraining. Additional details and results for the benchmarks and implementation can be found in the supplementary material, and we plan to make the source code publicly available.

Out-Of-Distribution Segmentation

In the context of OOD Segmentation, Cityscapes (Cordts et al. 2016) including 19 seen classes are used as the training sets, while OOD images containing other classes beyond the seen classes are utilized for testing purposes. Several approaches leverage OOD images with ground truth labels from larger datasets to enrich the training set, which overlaps with the OOD classes in the test set. Thus, to ensure fairness, all methods are differentiated based on the usage of OOD data and our proposed CSL is free of OOD data.

Comparison with SOTA Methods Table 1-4 show our results compared with existing methods on the SMIYC (Chan et al. 2021) Anomaly Track, Obstacle Track, LostAndFound-NoKnow (Pinggera et al. 2016), and Road Anomaly (Lis et al. 2019). There are 5 metrics for evaluation: (a) pixel-wise area under the precision-recall curve (AUPR), (b) pixel-wise false positive rate at a true positive rate of 95% (FPR_{95}), (c) adjusted Intersection over Union averaged over all ground truth segmentation components (sIoU gt), (d) component-wise F1-score averaged over different detection thresholds (mean F1), and (f) area under the receiver operating characteristics (AUROC). **SMIYC (Anomaly Track)** consists of real-world images, where each image may contain multiple OOD samples of different sizes from various categories. In SMIYC (AT), our proposed approach CSL outperforms all methods without OOD data by a substantial margin based on ObsNet, *e.g.*, CSL surpasses the former state of art method ObsNet (Besnier et al. 2021) by 4.7%, 19.5%, 2.3%, and 5.3% in AUPR, FPR_{95}, sIou gt, and mean F1. CSL even reaches state-of-the-art performance in terms of FPR_{95} and mean F1 across all methods including those leverage the OOD data. **SMIYC (Obstacle Track)** focuses on evaluating the ability to detect small-size obstacles on the road. In SMIYC (OT), CSL improves the former approach DaCUP (Vojíř and Matas 2023) by 5.6%, 0.43%, 7%, and 5% in AUPR, FPR_{95}, sIou gt, and mean F1 and achieve the state of art among all approaches in AUPR, FPR_{95}, and mean F1. **LostAndFound NoKnown** also focuses on evaluating the ability to detect small-size obstacles on the road and CSL improves the former approach DaCUP (Vojíř and Matas 2023) by 2.04%, 0.44%, 2.55%, and 0.12% in AUPR, FPR_{95}, sIou gt, and mean F1. And achieves state-of-the-art performance when combined with NFlowJS (Grcić, Bevandić, and Šegvić 2021). **Road Anomaly** has a similar setting with SMIYC (AT). As shown in Table 4, CSL achieves state of art performance among all methods including those with OOD data by improving the performance of ObsNet (Besnier et al. 2021) by 6.74%, 11.59%, and 4.3% in AP,

[1] Except for experiments marked with $*$, which uses ResNet100 as the backbone.

Method	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	motor	bicycle	mIoU
Source Only	79	39	75	26	25	34	34	39	82	18	84	58	37	70	19	15	5	22	54	43
+CSL$_2$	**82**	36	**78**	**28**	**29**	**40**	**45**	**48**	**83**	**25**	81	**68**	**45**	**81**	**24**	**20**	**7**	**24**	**57**	**47**$_{\uparrow 4}$
AdvEent	94	59	85	28	26	38	43	43	86	28	89	61	36	87	32	46	25	25	57	52
+CSL$_2$	**94**	**60**	**85**	28	**35**	**45**	**48**	**50**	**86**	28	**89**	**65**	**46**	**87**	**38**	**49**	**32**	24	**59**	**56**$_{\uparrow 4}$
DAFormer	96	73	89	40	44	49	53	60	58	49	91	71	45	91	75	77	64	55	61	65
+CSL$_2^*$	**96**	**74**	**90**	**51**	**48**	**52**	**56**	**65**	**89**	**48**	**91**	**76**	**45**	**93**	**77**	**80**	**68**	**56**	**66**	**70**$_{\uparrow 4}$

Table 6: Quantitative results for domain adaptation on the Synscapes2Cityscapes benchmark, where the source domain is the synthetic city scenes dataset (Synscapes) and the target domain is a real-world city scenes dataset (Cityscapes).

FPR_{95}, and AUROC.

Combination with Existing Methods without Retraining We combine CSL with three existing OOD segmentation methods (SML, Image Resynthesis, and ObsNet) with scheme 2, which doesn't require retraining, and compared their performance in Table 4. Noted that the performance of ObsNet is affected by the input size. Therefore, we use the ObsNet and ObsNet v2 to represent the experiments we use the original image size and fixed-smaller image size, *i.e.*, 512×1024. CSL outperformed all other methods in both benchmarks and even surpassed methods that use OOD data. Some methods achieved a decent AP but a poor FPR_{95} due to the difficulty of extracting OOD samples. ObsNet and ObsNet v2 achieved a high FPR_{95} in SMIYC (AT)-test, but our CSL significantly reduced it by 29.57 and 22.57, respectively. Figure 2 visually compares our proposed CSL with existing OOD segmentation methods, where we use ObsNet v2 to represent ObsNet due to its better performance. SML struggles to get acceptable results, and ObsNet produces decent AP but fails to achieve high recall with low FPR as shown in (e). In contrast, CSL demonstrates robustness to OOD samples as shown in (e).

Zero-Shot Semantic Segmentation

Table 5 presents a comparison of our proposed CSL method with previous state-of-the-art zero-shot semantic segmentation methods. We adopt the scheme_2 to integrate CSL with existing methods, primarily due to its reduced computational cost. (Section scheme_1 vs scheme_2). CSL outperforms ZegFormer, ZSSeg (Xu et al. 2022), and ZegCLIP by $0.9\%, 0.8\%, 0.2\%$ in seen classes, $3\%, 2\%, 1.4\%$ in unseen classes, and $2.1\%, 1.4\%, 0.8\%$ in harmonic classes in COCO-stuff benchmark and outperform those 3 methods by $5\%, 4.4\%, 1.6\%$ in unseen classes in PASCAL VOC 2012 benchmark. The experiment follows the same setting as ZegFormer, using 156 classes for training, and testing on all 171 classes from the COCO-stuff dataset.

Domain Adaptation in Semantic Segmentation

The CSL approach demonstrates superior performance not only on out-of-distribution (OOD) samples but also on in-distribution (ID) samples with domain gaps. Notably, our method achieves excellent results on the Synscapes2Cityscapes benchmark, as reported in Table 6. In these experiments, we use Synscapes, a synthetic city scene dataset, as the source domain, and Cityscapes, a real-world city scene dataset, as the target domain. And we also choose scheme_2 to integrate CSL with existing methods. CSL boost source-only by 4.02%, AdvEnt (Vu et al. 2019) by 3.5%, and DAFormer (Hoyer, Dai, and Van Gool 2022) by 4.1%.

Ablations

Negative Impact from Class Information Traditional methods for semantic segmentation assign each pixel from input images to prior semantic classes. However, this approach cannot handle OOD samples. The CA-RPG method assigns each pixel to N class-agnostic region prototypes, which learn more fundamental features that can represent both ID and OOD samples. Mask2Former and Zegformer also use a query-based framework, but introducing class supervision destroys the ability for OOD segmentation. The classification loss of ID classes causes the region prototypes to distribute within the subspace of ID classes, which makes it difficult to represent OOD classes effectively. In Figure 4, we can see the results of using the None-CA approach versus the CA-training approach on an image of a skier. The embeddings for the skier and background are not easily separable using the None-CA approach, while the CA-training approach allows for a better representation of both classes.

CA Training and Soft Assignment Quantitative results in Table 7 demonstrate the effectiveness of CA training and our proposed soft assignment (SA). Before evaluation, we count the ground truth labels corresponding to all pixels in each region proposal and select the label with the highest frequency as the class of the entire region. This post-processing method is proposed in SMIYC (Chan et al. 2021) and allows us to use the same evaluation criteria (mIoU, fwIoU, mACC, and pACC) as semantic segmentation to assess the quality of region proposals.

We present a comparison of three approaches for training a region proposal generator: None-CA, which employs binary mask and classification loss; CA+HA, which employs CA training and hard assignment during inference; and CA+SA, our proposed approach which combines CA training with soft assignment during inference. The model is trained on Cityscapes and tested on COCO-stuff, where ID represents seen classes from Cityscapes, and OOD represents those in the COCO but not in Cityscapes. Results

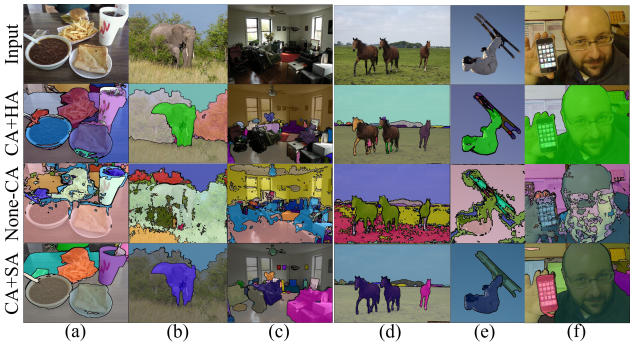

Figure 3: Visualisations of the efficacy of CA-training and soft assignment. CA+HA represents CA-M2F, where the model is trained in a class-agnostic way and inferences via the hard assignment, None-CA represents the model is trained with class loss and inferences via the soft assignment, and CA+SA represents CSL, where the model trained in a class-agnostic way and inferences via the soft assignment. Note that the model is only trained on the Cityscapes and tested on COCO-stuff.

	None-CA	CA+HA	CA+SA
mIoU	45.99	$49.17_{\uparrow 3.18}$	$51.34_{\uparrow 5.35}$
ID-mIoU	68.44	$76.35_{\uparrow 7.91}$	$77.20_{\uparrow 8.76}$
OOD-mIoU	43.83	$46.56_{\uparrow 2.73}$	$48.85_{\uparrow 5.02}$
mACC	58.46	$60.31_{\uparrow 1.85}$	$63.10_{\uparrow 4.64}$

Table 7: Ablation study of CA-training and Soft Assignment. The model is trained on the Cityscapes-train and tested on the COCO-stuff.

in Table 7 demonstrate that both CA training and soft assignment significantly improve performance across all metrics. Figure 3 visually illustrates the improvement. None-CA fails on most unseen objects, while CA+HA produces decent results but struggles with challenging cases such as indoor scenes, multiple animals, and small accessories. Soft assignment overcomes the limitations of hard assignment by assigning regions pixel-wise, providing more refined segmentation results.

Comparison with Segment Anything Model A notable contribution of CSL is its capability to segment out-of-distribution (OOD) objects without relying on any OOD data, utilizing only minimal training data. In this section, we employ a foundation model, SAM (Kirillov et al. 2023), which is used by many recent works (Zhang, Li, and Ahuja 2023), to produce CA region proposals and subsequently integrate it with ObeNet (Besnier et al. 2021), DACUP (Vojíř and Matas 2023), and NFlowJS (Grcić, Bevandić, and Šegvić 2021) in the SMIYC-AT, OT, and LAF NoKnown benchmarks. This approach yields an improvement of 1.7% in AUPR and 0.3% in FPR_{95} on average across those three benchmarks compared with CSL in Table 1, 2, and 3, which demonstrates that the quality of CA region proposals generated by CSL is satisfied, even in the absence of any OOD data. We believe the constraining factor influencing the outcome appears to be the classification accuracy of each region, rather than segmentation quality. More results are shown in the Appendix.

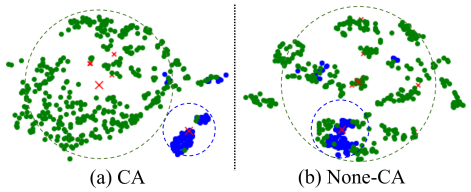

Figure 4: Embedding visualisations of Figure 3-(e) by T-SEN. We plot the region prototypes as red times symbols, the per-pixel embeddings from the background as green bullets, and the skier as blue bullets. The sizes of the prototypes indicate the validity scores.

Scheme_1 vs Scheme_2

In Tables 1, 2, and 3, we present results for scheme 1 and 2. While scheme_1 trails by approximately 0.3% in AUPR relative to scheme_2, the results on FPR_{95} display a mix of advantages for both methods. Notably, scheme_2 demonstrates efficiency in training, requiring half the iterations to match the performance of scheme_1. For context, in our integration experiments with ZegCLIP (Zhou et al. 2023) on the COCO-stuff benchmark, scheme_1 demanded around 50K iterations to achieve satisfactory results, whereas scheme_2 reached similar benchmarks in just 25K iterations. However, another key consideration is the inference time. Scheme_1, being an end-to-end solution, is more efficient during inference: in our evaluations, scheme_2 took 33% longer on average across all conducted experiments."

Conclusion

This paper presents the Class-Agnostic Structure-Constrained Learning (CSL) method for addressing the challenge of segmenting the unseen. CSL provides 2 different schemes, which can be utilized as an end-to-end framework or integrated with existing methods without retraining. Our experimental results demonstrate that CSL outperforms existing state-of-the-art methods across 3 challenging tasks. Moreover, we have provided an analysis of the reasons behind the effectiveness of our proposed method. We believe that the ability of CSL to learn about classes not seen during training, by eliciting class-agnostic information from the ID images, is a crucial factor contributing to its superior performance. Overall, CSL provides a promising solution for segmenting the unseen, and we hope our work will lead to other related work in this area.

Acknowledgements

The support of the Office of Naval Research under grant N00014-20-1-2444 and of USDA National Institute of Food and Agriculture under grant 2020-67021-32799/1024178 are gratefully acknowledged.

References

Baur, C.; Wiestler, B.; Albarqouni, S.; and Navab, N. 2019. Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, 161–169. Springer.

Besnier, V.; Bursuc, A.; Picard, D.; and Briot, A. 2021. Triggering Failures: Out-Of-Distribution detection by learning from local adversarial attacks in Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15701–15710.

Bevandić, P.; Krešo, I.; Oršić, M.; and Šegvić, S. 2018. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*.

Bevandić, P.; Krešo, I.; Oršić, M.; and Šegvić, S. 2019. 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*, 33–47. Springer.

Chan, R.; Lis, K.; Uhlemeyer, S.; Blum, H.; Honari, S.; Siegwart, R.; Fua, P.; Salzmann, M.; and Rottmann, M. 2021. Segmentmeifyoucan: A benchmark for anomaly segmentation. *arXiv preprint arXiv:2104.14812*.

Chan, R.; Rottmann, M.; and Gottschalk, H. 2021a. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5128–5137.

Chan, R.; Rottmann, M.; and Gottschalk, H. 2021b. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5128–5137.

Cheng, B.; Choudhuri, A.; Misra, I.; Kirillov, A.; Girdhar, R.; and Schwing, A. G. 2021. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 3213–3223.

Creusot, C.; and Munawar, A. 2015. Real-time small obstacle detection on highways using compressive RBM road reconstruction. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, 162–167.

Di Biase, G.; Blum, H.; Siegwart, R.; and Cadena, C. 2021. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16918–16927.

Ding, J.; Xue, N.; Xia, G.-S.; and Dai, D. 2022. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11583–11592.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Grcić, M.; Bevandić, P.; and Šegvić, S. 2021. Dense anomaly detection by robust learning on synthetic negative data. *arXiv preprint arXiv:2112.12833*.

Grcić, M.; Šarić, J.; and Šegvić, S. 2023. On Advantages of Mask-level Recognition for Outlier-aware Segmentation. arXiv:2301.03407.

Haldimann, D.; Blum, H.; Siegwart, R.; and Cadena, C. 2019. This is not what i imagined: Error detection for semantic segmentation through visual dissimilarity. *arXiv preprint arXiv:1909.00676*.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2018a. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2018b. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*.

Hoyer, L.; Dai, D.; and Van Gool, L. 2022. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9924–9935.

Jung, S.; Lee, J.; Gwak, D.; Choi, S.; and Choo, J. 2021. 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*, 15425–15434.

Kang, B.; Kwak, J.; and Kang, S.-J. 2022. Anomaly Segmentation Using Class-aware Erosion and Smoothing. In *2022 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, 1–4.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information processing Systems*, 30.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30.

Lee, K.; Lee, H.; Lee, K.; and Shin, J. 2017. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*.

Liang, S.; Li, Y.; and Srikant, R. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft

coco: Common objects in context. In *European Conference on Computer Vision*, 740–755.

Lis, K.; Honari, S.; Fua, P.; and Salzmann, M. 2020. Detecting road obstacles by erasing them. *arXiv preprint arXiv:2012.13633*.

Lis, K.; Nakka, K.; Fua, P.; and Salzmann, M. 2019. Detecting the unexpected via image resynthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2152–2161.

Liu, Y.; Ding, C.; Tian, Y.; Pang, G.; Belagiannis, V.; Reid, I.; and Carneiro, G. 2022. Residual Pattern Learning for Pixel-wise Out-of-Distribution Detection in Semantic Segmentation. *arXiv preprint arXiv:2211.14512*.

Liu, Y.; Tian, Y.; Maicas, G.; Pu, L. Z. C. T.; Singh, R.; Verjans, J. W.; and Carneiro, G. 2020. Photoshopping colonoscopy video frames. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 1–5.

Ma, X.; Li, B.; Wang, Y.; Erfani, S. M.; Wijewickrema, S.; Schoenebeck, G.; Song, D.; Houle, M. E.; and Bailey, J. 2018. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*.

Mukhoti, J.; and Gal, Y. 2018. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*.

Nayal, N.; Yavuz, M.; Henriques, J. F.; and Güney, F. 2023. RbA: Segmenting Unknown Regions Rejected by All. arXiv:2211.14293.

Ohgushi, T.; Horiguchi, K.; and Yamanaka, M. 2020. Road obstacle detection method based on an autoencoder with semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision*.

Pinggera, P.; Ramos, S.; Gehrig, S.; Franke, U.; Rother, C.; and Mester, R. 2016. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1099–1106.

Qi, L.; Kuen, J.; Wang, Y.; Gu, J.; Zhao, H.; Torr, P.; Lin, Z.; and Jia, J. 2022. Open World Entity Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Tian, Y.; Liu, Y.; Pang, G.; Liu, F.; Chen, Y.; and Carneiro, G. 2022. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *European Conference on Computer Vision*, 246–263.

Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J. W.; and Carneiro, G. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4975–4986.

Vandenhende, S.; Georgoulis, S.; Proesmans, M.; Dai, D.; and Van Gool, L. 2020. Revisiting multi-task learning in the deep learning era. *arXiv preprint arXiv:2004.13379*, 2(3).

Vojíř, T.; and Matas, J. 2023. Image-Consistent Detection of Road Anomalies As Unpredictable Patches. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5491–5500.

Vojir, T.; Šipka, T.; Aljundi, R.; Chumerin, N.; Reino, D. O.; and Matas, J. 2021. Road anomaly detection by partial image reconstruction with segmentation coupling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15651–15660.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2517–2526.

Xia, Y.; Zhang, Y.; Liu, F.; Shen, W.; and Yuille, A. L. 2020. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *European Conference on Computer Vision*, 145–161. Springer.

Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; and Bai, X. 2022. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, 736–753. Springer.

Zhang, H.; Li, F.; and Ahuja, N. 2023. Open-NeRF: Towards Open Vocabulary NeRF Decomposition. *arXiv preprint arXiv:2310.16383*.

Zhang, H.; and Zhang, R. 2022. Active domain adaptation with multi-level contrastive units for semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 1640–1657.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3): 302–321.

Zhou, Z.; Lei, Y.; Zhang, B.; Liu, L.; and Liu, Y. 2023. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11175–11185.