2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Unsupervised Road Anomaly Detection with Language Anchors[†]

Beiwen Tian[1,2], Mingdao Liu[1,2], Huan-ang Gao[1,2], Pengfei Li[1,2], Hao Zhao[2] and Guyue Zhou[2,*]

***Abstract*— Road anomaly detection is critical to safe autonomous driving, because current road scene understanding models are usually trained in a closed-set manner and fail to identify unknown objects. What's worse, it is difficult, if not impossible, to collect a large-scale dataset with anomaly annotations. So this paper studies unsupervised anomaly detection which finds out anomaly regions using scene parsing logits solely. While former methods depend on the weights learned from the closed training set as anchors for logit generation, we resort to language anchors that are learned from enormous paired vision and language data. Thanks to rich open-set semantic information contained in these language anchors, our method performs better than former unsupervised counterparts while maintaining the advantage of training without accessing any out-of-distribution data. We delve into this new paradigm and identify the superiority of using pairwise binary logits, which we credit to a better understanding of the negation language anchor. Last but not least, we find that the former top-1 selection of semantic labels for uncertainty measurement is problematic in many cases and a new blended standardization strategy brings clear improvements to our solution. We report state-of-the-art performance on FS LostAndFound, LostAndFound and RoadAnomaly datasets among comparable methods. The codes are publicly available at https://github.com/TB5z035/URAD-LA.git**

I. Introduction

Modern learning-based vision algorithms have tremendously boosted the environment sensing capabilities of autonomous robots in diverse applications [1]–[6]. However, semantic understanding [7]–[10], which is a fundamental robotic vision module, has long been addressed in a closed-set manner that assumes a fixed number of known semantic categories. Models trained this way usually fail to identify samples that do not fall into any of the pre-defined classes and produce false predictions silently [11]. Therefore, recently the robotics community has proposed many methods to equip vision models with anomaly detection ability, especially for safety-critical scenarios including autonomous driving [12], drone delivery [13], minimally-invasive surgery [14], and many others [15]–[19], where flexibility and robustness to unexpected input is indispensable.

Some existing works strive to identify anomalous objects with labelled out-of-distribution (OoD) samples, which are collected from real-world application scenarios [20] or by injecting artificial anomalies into regular scenes [21]. However, unexpected inputs are typically rare and sparsely distributed in the application scenarios, making it difficult to collect and annotate OoD data with sufficient quantity and diversity. Such limited OoD samples may only help the models detect *known anomalies* but cannot guarantee the anomaly detection ability to generalize to real-world unexpected situations.

Unsupervised anomaly detection [13], [14], [22], [23], on the other hand, aims to assess model uncertainty directly from a pre-trained vision model. This paradigm does not require OoD data or extra training and is expected to reflect the perception uncertainty intrinsic to the model. However, there still exists a large performance gap between unsupervised and supervised anomaly detection methods. We hypothesize that a critical factor in their limited performance is that all the information they use to identify anomalies still comes from a closed training set. And introducing external open-set knowledge is a potential solution.

To this end, we introduce vision-language pretraining (VLP) into unsupervised anomaly detection and find that it works surprisingly well. We focus on the well-benchmarked semantic segmentation task for urban driving scenes, which is a challenging case with complex visual inputs. As shown in Fig. 1, our approach significantly suppresses difficult false positives caused by tiles (row 1) and graffiti (row 2), and successfully identifies large (row 3) and faraway (row 4) anomaly regions. By contrast, the performances of former state-of-the-art (SOTA) is unsatisfactory in these scenarios.

As conclusively shown in the later Table. III, the efficacy of our approach is mainly attributed to the transferred open-set knowledge from vision-language pretraing. To leverage the strong embedding space from VLP, we substitute logit anchors obtained from closed-set supervision, with logit anchors generated by vision-language pretraining. This provides a more robust and transferable feature space that better separates the representation of inliers (i.e., in-distribution samples) from that of the outliers (i.e., out-of-distribution samples). Within this new framework, we identify several key technical practices: (1) we adopt binary logits to help the image encoder learn better representations from the negation language anchor. (2) we present a novel blended logit standardization strategy that addresses critical issues of top-1 semantic category assignment in uncertainty measurement. (3) we show the widely used practice of multi-scale inference also works for anomaly detection.

The effectiveness of our approach is demonstrated by the state-of-the-art performance on RoadAnomaly, LostAndFound and Fishyscapes LostAndFound, outperforming previous comparable SOTA [22] by 7.51%-22.90% improvement in AP and 2.61%-5.17% reduction in FPR_{95}. Our contributions can be summarized as follows:

[†]This work is funded by DiDi GAIA Initiative
[*]Corresponding author
[1]Department of Computer Science and Technology, Tsinghua University, China, {tbw18,liu-md20,gha20,li-pf22}@mails.tsinghua.edu.cn.
[2]Institute for AI Industry Research (AIR), Tsinghua University, China, {zhaohao, zhouguyue}@air.tsinghua.edu.cn.

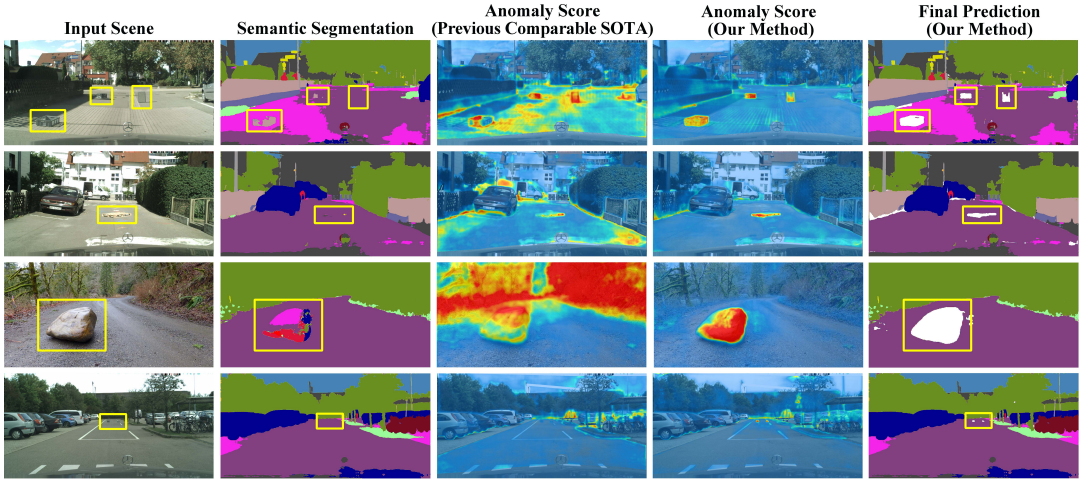

Fig. 1. **Qualitative results compared with previous comparable SOTA SML [22].** Anomalous objects in the **Input Scene** are highlighted in yellow rectangles, and the **Semantic Segmentation** shows closed-set results where each pixel is classified into one of the pre-defined classes. Our method can better address complex scenarios, such as multiple anomalies, road with graffiti and forest lanes. Our method produces accurate anomaly segmentation with significantly fewer false positives (rows 1-3) and detects faraway anomalies (row 4). Color closer to red denotes a higher estimated probability to be anomalous. Zoom in for a better view.

- We propose the first unsupervised road anomaly detection framework that exploits vision-language pretraining for logit anchoring.
- We identify the importance of using pairwise binary logits to better exploit the negation language anchor.
- We show the drawback of top-1 category assignment and propose a blended logit standardization strategy.
- We achieve SOTA performance on RoadAnomaly, LostAndFound, and Fishyscapes LostAndFound among comperable methods and release our code.

II. Related Works

Anomaly detection in urban driving scenes. Anomaly detection aims to identify outliers from inliers, and is typically performed through a scoring function that assigns an anomaly score to each pixel of the input image. Many road anomaly detection methods depend on OoD data such as the void class in Cityscapes training set [24], or images from COCO [21] or ImageNet [25], which are pasted into normal driving scenes as anomalous objects. Another stream of works [24], [26]–[28] detect anomalous regions through image re-synthesis, segmenting the anomaly by discrepancies between the input and the generated image.

Most unsupervised approaches assume that anomalous samples yield higher uncertainty than inliers. For uncertainty measurement, some delve into Bayesian estimations [29], [30], like MC Dropout [31] and Contrete Dropout [32], while others resort to the logits from the final layer, like Max Softmax Probability (MSP) [23], Entropy [23] and Max Logits [33]. The logit-based method is further improved by Standardized Max Logit (SML) [22], which normalizes the max logit with statistics according to the class prediction.

Vision-Language Pretraining. Aiming at associating vision and language concepts, vision-language pretraining has made impressive progress in recent years [34]–[36]. As a milestone, Contrastive Language-Image Pretraining (CLIP) [36] achieves remarkable performance in zero-shot image classification by learning multi-modal representations from enormous text-image pairs on the Internet. Though trained on image level, CLIP also shows the potential to boost dense prediction tasks, including object-detection [37], referring image segmentation [38] and semantic segmentation [39].

III. Method

In this section, we introduce our unsupervised approach to help autonomous robots better detect anomalies, which is applicable to various robotic scenarios. Specifically, we consider its application on urban autonomous driving, in which we detect anomalous objects in the input road scene. We first describe our procedures to obtain language-based binary logits for each class along with the negation anchor (See Sec. III-A). Then, we put forward a novel strategy for uncertainty measurement using the blended logits (See Sec. III-B). An overview of our method is shown in Fig. 2.

A. Language-based Logits

Overview. While supervised anomaly detection methods typically have access to different extents of outlier exposure, previous unsupervised methods learn from a closed training set only, resulting in the significantly lower benchmark results. The lack of the open-set prior is the key problem of these arts, which we believe can be alleviated by VLP (e.g. CLIP [36]). Therefore, we exploit the feature space of VLP models, in which we embed the input image and adopt language anchors to produce logits for classification. We further adopt binary logits with the negation language anchor, which builds a better-preserved representation space from the pairwise anchor embeddings.

Encoding of Language Anchors. Suppose we have K classes in the training set for semantic segmentation and a C-dimensional feature space from the vision-language pretraining. We feed the class labels (i.e. the names of K

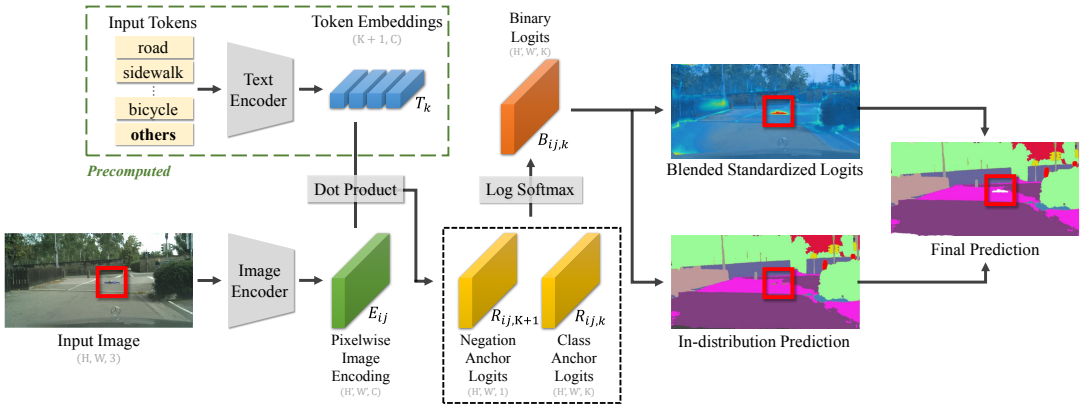

Fig. 2. **Overview of our method.** Suppose there are K pre-defined classes in the training set. We produce $(K+1)$ text embeddings, corresponding to K class anchors and 1 negation anchor (i.e. the embedding of token *others*). The text embeddings and pixel-wise image encoding are correlated with a dot product, producing pixel-wise logits for each class anchor ($R_{ij,k}$) and the negation language anchor ($R_{ij,K+1}$). Then, for each class k, we conduct a log-softmax between its class anchor logit and the negation anchor logit which yields the binary logit $B_{ij,k}$. Blended Standardized Logits are produced from $B_{ij,k}$ and used as the anomaly score for final prediction.

classes) into the text encoder, producing K language anchors $T_1, \cdots, T_K \in \mathbb{R}^C$. We also feed the token *others* into the text encoder to produce the negation language anchor $T_{K+1} \in \mathbb{R}^C$. The $(K+1)$ anchors are pre-computed and fixed in the subsequent training process. This process is demonstrated in the green dotted box in Fig. 2.

Encoding of Image Features. The image encoder generates pixel-wise feature encodings for the down-sampled image. Suppose the input RGB image has the shape $(H, W, 3)$, then the image encoder yields a feature map $E \in \mathbb{R}^{H' \times W' \times C}$ (illustrated by green volumes in Fig. 2) with the height $H' = \frac{H}{4}$ and the width $W' = \frac{W}{4}$.In training, the image encoder learns to produce feature encodings in the same representation space as the language anchors.

Anchor Logits. We exploit the common feature space by correlating visual and text features with dot product. In this process, we denote the encoded feature at pixel (i, j) by E_{ij} and define the anchor logit of pixel (i, j) w.r.t. class k as:

$$R_{ij,k} = E_{ij} \cdot T_k \tag{1}$$

Combining the logits from all K classes, we define the similarity between pixel (i, j) and the pre-defined classes as $R_{ij} = (R_{ij,1}, \cdots, R_{ij,K})$. Similarly, we correlate the negation anchor with the pixel-wise feature encodings to produce $R_{ij,K+1}$. In this way, we obtain the pixel-wise class anchor logits for the image with the shape of (H', W', K) along with the negation anchor for the image with the shape of $(H', W', 1)$ (illustrated by yellow volumes in Fig. 2).

Binary Logits. We leverage the negation language anchor to perform binary classification, the results of which are taken as the final logits. More specifically, we define the binary logit of pixel (i, j) w.r.t. class k as the logarithm of the softmax probability to binarily classify this pixel into class k against class *others*, namely:

$$B_{ij,k} = \log\left(\frac{\exp(R_{ij,k}/\tau_1)}{\exp(R_{ij,k}/\tau_1) + \exp(R_{ij,K+1}/\tau_1)}\right) \tag{2}$$

where τ_1 is a hyper-parameter of temperature. The binary logits are crucial to the representation transfer and the ability of generalizing to anomalous samples, as is manifested in later Table. II. Combining the results of all pixels on the K classes, we obtain the binary logits for the whole image $B \in \mathbb{R}^{H' \times W' \times K}$ (orange volumes in Fig. 2), which is then up-sampled to the original resolution of the input as (H, W, K).

B. Blended Standardized Logits

With the language-based binary logits $B_{ij,k}$, we propose a novel measurement in this stage to calculate the pixel-wise uncertainty (or the anomaly score) by standardization of logits. Logit standardization is a technique first proposed by a SOTA method Standardized Max Logits (SML) [22], which collects and standardizes the logits in a class-wise manner for all pixels to obtain the anomaly scores so that no OoD data or retraining is needed. Before we formally introduce our proposed Blended Standardized Logits (BSL), we first describe the formal procedures for logit standardization.

Standardized Logits. To perform logit standardization, we first obtain the predicted category $\hat{Y}_{ij}$ and the max-logit M_{ij} for each pixel by:

$$\hat{Y}_{ij} = \arg\max_k \ B_{ij,k} \tag{3}$$

$$M_{ij} = \max_k \ B_{ij,k} \tag{4}$$

Next, the mean and variance of max-logits for each class k are estimated from samples in the training set:

$$\mu_k = \frac{\sum_i \sum_{h,w} \mathbb{1}\left(\hat{Y}_{hw}^{(i)} = k\right) \cdot M_{hw}^{(i)}}{\sum_i \sum_{h,w} \mathbb{1}\left(\hat{Y}_{hw}^{(i)} = k\right)} \tag{5}$$

$$\sigma_k^2 = \frac{\sum_i \sum_{h,w} \mathbb{1}\left(\hat{Y}_{hw}^{(i)} = k\right) \cdot \left(M_{hw}^{(i)} - \mu_k\right)^2}{\sum_i \sum_{h,w} \mathbb{1}\left(\hat{Y}_{hw}^{(i)} = k\right)} \tag{6}$$

where $\mathbb{1}(\cdot)$ is the indicator function and (i) represents the i-th sample in the training set. Then in a class-wise manner, the logits are standardized with these statistics:

$$S_{ij,k} = \frac{B_{ij,k} - \mu_k}{\sigma_k} \tag{7}$$

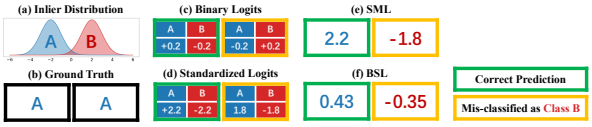

Fig. 3. **Motivation of BSL.** Suppose we have Class A and B with inlier distribution $\mu_A = -2, \mu_B = 2, \sigma_A^2 = \sigma_B^2 = 1$ before standardization in (a). Now consider the classification of **two** pixels with ground truth (b) of class A. The produced binary logits (c) are similar for the two classes but standardized logits (d) are quite different. As a result, Standardized Max Logits (e) yield anomaly scores with a drastic difference, which can lead to serious false positives. Blended Standardized Logits (f), on the other hand, smooth the anomaly score on pixels with higher uncertainty and alleviate false positives due to misclassification.

Finally, in SML [22], the anomaly score for pixel (i, j) is defined as:

$$A_{ij}^{(\text{SML})} = \sum_k \mathbb{1}(\hat{Y}_{ij} = k) S_{ij,k} \tag{8}$$

Since the logit distributions vary significantly among classes, the standardization makes them more comparable so that the outliers are better separated. The anomaly score $A_{ij}^{(\text{SML})}$ calculated in SML [22] is assumed to subject to a standard Gaussian for inliers and deviate from zero for outliers.

Blended Standardized Logits. Despite the simplicity of this paradigm, we argue that critical flaws exist in the definition of anomaly scores proposed by SML [22]. As shown in the example of Fig. 3, when an inlier is misclassified (illustrated as the yellow box), its Standardized Max Logit may differ drastically from those of pixels that have similar binary logits. We observe that the discontinuous way of choosing the logits (i.e. always taking the maximum logits only) leads to more false positives in the misclassified region. We also notice that many of the misclassified pixels have softmax probabilities of the same magnitude, which may be exploited to mitigate such inaccuracy.

To address the aforementioned issue, we propose a novel Blended Standardized Logit, which aims to measure the uncertainty of a pixel by the weighted average of all possible standardized logits instead of that of the predicted class only. The BSL of pixel (i, j), i.e. $A_{ij}^{(\text{BSL})}$, is defined as the weighted sum that blends all standardized logits with softmax probability as weights:

$$A_{ij}^{(\text{BSL})} = \sum_k \frac{\exp(B_{ij,k}/\tau_2)}{\sum_m \exp(B_{ij,m}/\tau_2)} S_{ij,k} \tag{9}$$

where τ_2 is a hyper-parameter for temperature, and we take $A_{ij}^{(\text{BSL})}$ as the anomaly score for pixel (i, j). Surprisingly, as is shown in Fig. 7, BSL not only reduces false positives for inliers but also decreases false negatives on anomalies, resulting in significant improvement in the performance of anomaly detection (reported in Table. I).

C. Multi-scale Inference

Most anomaly detection approaches simply take this anomaly score map as the final prediction. However, as convolution layers are not scale-invariant, objects that are too large or too small in the image may be misidentified. For instance, an inlier posed too close to the camera tends to be recognized as anomaly whereas an outlier placed too far away mixes up with the background and tends to be predicted as the same class with high confidence.

To address this issue that occurs on particular scales, we adopt multi-scale inference, which is a post-processing technique widely-used in semantic segmentation [37] and object detection [40] but barely visited in anomaly detection. At inference time, the input scene is resized to different scales to produce respective anomaly score maps. These score maps are then interpolated to the original resolution then averaged to generate the final anomaly score map.

IV. Experiment

A. Datasets and Implementation Details

Datasets. (1) **LostAndFound** [41] is one of the first public datasets for anomaly detection, with diverse real-world anomalous objects in urban driving scenes. The LostAndFound test set provides 1203 images captured in 13 street scenes featuring 37 anomaly types, and contains challenging scenarios including distant anomalous objects, various road surfaces, and illumination shifts. Note that only pixels of road regions and anomalous objects are annotated in the LostAndFound. (2) **Fishyscapes benchmark** [42] provides high-resolution images for anomaly detection in urban driving scenes. Within this benchmark, Fishyscapes LostAndFound is a subset of the LostAndFound dataset with additional semantic annotations for all pixels. The validation set of Fishyscapes LostAndFound is publicly accessible and contains 100 images in total. (3) **RoadAnomaly** [26] is another dataset for anomaly detection and contains 60 online-crawled images with various unexpected objects annotated in a pixel-wise manner. RoadAnomaly is especially challenging since the anomaly objects are of various scales.

Implementation Details. In our method, the image encoder is implemented by DeepLabv3 [43] with ResNet101 [44] backbone. For language anchors, CLIP [36] (ViT-B32 [45] variant) is employed as the text encoder.

To train the network, we first feed the class labels into the text encoder to obtain the language anchors which are fixed during training. The network is then trained on the task of semantic segmentation on Cityscapes training set with AdamW [46] optimizer. Before feeding the image samples to the image encoder, we apply stochastic transformations including random cropping, scaling, Gaussian blur, and color jitters. We select separate checkpoints for each dataset and report the best result.

Upon inference time, we adopt dilated kernel smoothing technique following [22] without boundary suppression. The inference scales are a combination of smaller (0.5, 0.65, 0.85), original (1.0), and larger (1.25, 1.75). Their effects are further discussed in the ablation study.

Evaluation Details. For quantitative results, adopted metrics include area under receiver operating characteristics (AUROC), average precision (AP), and the false positive rate at a true positive rate of 95% (FPR_{95}). For qualitative results, since the ranges of anomaly score vary between models and images, we generate the visualization heatmap

TABLE I

COMPARISONS ON ANOMALY DETECTION BENCHMARKS

Anchor	Method	FS LostAndFound			RoadAnomaly			LostAndFound		
		AUROC ↑	AP ↑	FPR_{95} ↓	AUROC ↑	AP ↑	FPR_{95} ↓	AUROC ↑	AP ↑	FPR_{95} ↓
Closed-set	MSP	88.17	4.25	44.60	74.36	21.48	65.46	93.71	29.83	27.42
	Entropy	89.57	10.72	43.50	76.43	23.96	64.24	94.99	50.29	25.52
	Max Logit	92.13	17.99	39.76	79.73	26.56	59.04	**97.21**	**65.71**	**15.22**
	SML [22]	**97.24**	**38.09**	**14.71**	**82.86**	**28.06**	**48.55**	95.19	56.30	24.97
Language-based (Ours)	MSP	90.39	3.91	35.08	73.75	20.17	64.16	95.20	32.76	19.18
	Entropy	92.18	7.61	34.52	75.24	22.15	63.28	96.74	53.03	16.65
	Max Logit	96.24	41.94	23.62	81.23	29.41	58.43	97.65	66.75	13.70
	SML [22]	97.83	54.93	11.70	87.35	48.37	48.96	97.20	65.88	16.67
	BSL (Ours)	**98.08**	**60.99**	**10.62**	**88.26**	**48.73**	**45.94**	**98.26**	**73.22**	**10.05**

(e.g. Fig. 4(a)(b)) with anomaly score thresholds from TPR_{20} to FPR_{80} in respective datasets, so that the visualizations are comparable between methods. Color closer to red denotes a higher estimated probability to be anomalous.

B. Benchmark Results

We evaluate our method and former unsupervised baselines with the closed-set and language-based anchors, and report the performances in Table I. Since our method requires neither OoD data nor extra network and reuses the segmentation network without re-training, the unsupervised baselines for comparison are chosen by the same criteria.

Compared with former SOTA (closed-set SML [22]), the proposed language-based BSL significantly improves AP by 22.9% and reduces FPR_{95} by 4.09% on the FS LostAndFound validation set. On the RoadAnomaly dataset, our method also improves AP by 20.67% and decreases FPR_{95} by 2.61%. Given the fact that most anomalies in the FS LostAndFound validation set are tiny objects in various shapes and distances, and the fact that anomalies have various scales and styles in the Road Anomaly dataset, our superior performances on both datasets manifest the robustness of our method to the shapes and scales of anomalies.

Another intriguing observation is that SML [22] is outperformed by the simpler Max Logit baseline on the LostAndFound dataset. We attribute this phenomenon to the fact that logits of the misclassified regions are standardized by statistics of the wrong categories. Considering that *road* is the only inlier category in the LostAndFound dataset, the misclassified category is entirely unreliable and using the statistics of the wrong category for standardization leads to drastic performance drop. By contrast, our BSL method naturally combines the logits with probability-based weights and produces higher performance than SML and other baselines.

C. Ablation Study

Binary Logits with Negation Anchor. We conduct experiments on the Fishyscapes LostAndFound validation set using the class anchor logits $R_{ij,k}$ and the binary logits $B_{ij,k}$ (see Fig. 2) respectively, while the other settings are identical. Results reported in Table II demonstrate that using binary logits brings significant improvement in all three metrics over

TABLE II

COMPARISON ON DIFFERENCE LOGIT TYPES

Method	AUROC ↑	AP ↑	FPR_{95} ↓
Anchor Logits	93.17	13.57	29.44
Binary Logits	**98.08**	**60.99**	**10.62**

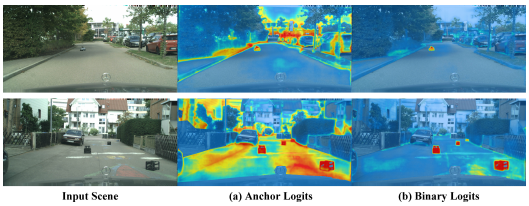

Fig. 4. **Qualitative comparisons on different logit types.** Compared with binary logits, anchor logits typically fail on (1) boundaries between known classes, (2) hard samples, like buildings in the distance (row 1) and road region with graffiti (row 2). Zoom in for a better view.

the class anchor logits. Visualization in Fig. 4 also shows that using binary logits results in significantly less false positives.

We attribute the superior perforance of binary logits to a better-transferred representation space with the effective utilization of the negation language anchor. During training for semantic segmentation, the negation language brings about *extra supervision* so that the image encoder is enforced to maximize the distance between the pixel-wise encoding and the negation language anchor as well, leading to a better preserved structure of the feature space. We provide proof by the distributions of anomaly scores using different logit types (depicted in Fig. 5) and show that the predicted inliers have similar distributions whereas the outliers are better separated

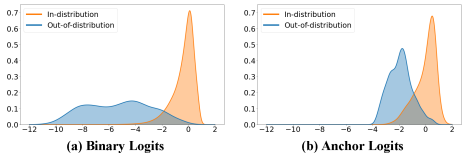

Fig. 5. **Probability density of different logits types.** The numerical ranges of axes are aligned. While in-distribution samples are clustered similarly, binary logits can better separate the anomalous samples, resulting in better anomaly detection performance.

TABLE III

AVERAGE METRIC GAINS WITH LANGUAGE ANCHORS

Method	AUROC ↑	AP ↑	FPR_{95} ↓
MSP	1.03	0.43	-6.35
Entropy	1.06	-0.73	-6.27
Max Logit	2.02	9.28	-6.09
SML	2.36	15.58	-3.63
BSL	3.10	18.91	-7.57

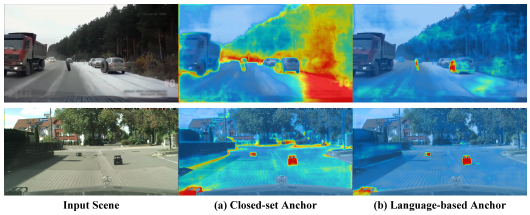

Fig. 6. **Qualitative comparisons on different anchor types.** Compared with closed-set anchors (a), language anchors (b) can generalize better to unseen scenes with drastic domain shifts (row 1) and reduce false positives produced by hard samples and boundaries (row 2).

in the feature spaces learned with binary logits.

Language Anchors. In Table III we report the average metric gains on the three datasets after applying language-based anchors. Using language anchors significantly improves AP for logits-based methods and reduces FPR_{95} for all methods. We further provide visualization of anomaly scores obtained using language anchors and closed-set anchors in Fig. 6. Language anchors precisely segment the anomalies in scenes that largely differ from the training set we use (row 1). Language anchors also reduce false positives caused by hard samples and boundaries (row 2), which is a pitfall for the previous unsupervised approaches.

Blended Standardized Logits (BSL). The last two rows of Table I manifest the effect of BSL on the three datasets. Compared with SML, BSL improves AP by 0.36%-7.34% and decreases FPR_{95} by 1.08%-6.62% using the same anchor and logit type. We provide qualitative comparisons between BSL and SML in Fig. 7 and show that BSL can better identify large anomalies (row 1), reduce false positives (row 2) and distinguish distant anomalous objects (row 3). This shows that BSL is a better measurement for uncertainty.

Multi-scale Inference. Table IV shows the influence of different scale combinations on the Fishyscapes LostAndFound validation set. Multi-scale inference significantly reduces FPR_{95} while retaining acceptable AP. The anomaly scores obtained by different inference scales are compared in Fig. 8. Larger scales help identify the distant anomalies that are hard to distinguish from the background (row 2), while smaller scales have the *smoothing* effect which reduces false positives of hard examples with confusing texture (e.g. tiled sidewalk in row 1). Hence, a combination of various scales yields a more robust performance.

TABLE IV

COMPARISON ON DIFFERENT INFERENCE SCALES

Smaller	Original	Larger	AUROC ↑	AP ↑	FPR_{95} ↓
✓			97.18	61.91	17.73
	✓		97.39	51.44	15.05
		✓	95.48	29.02	24.52
✓	✓		97.56	**63.94**	14.78
	✓	✓	96.98	43.25	16.10
✓	✓	✓	**98.08**	60.99	**10.62**

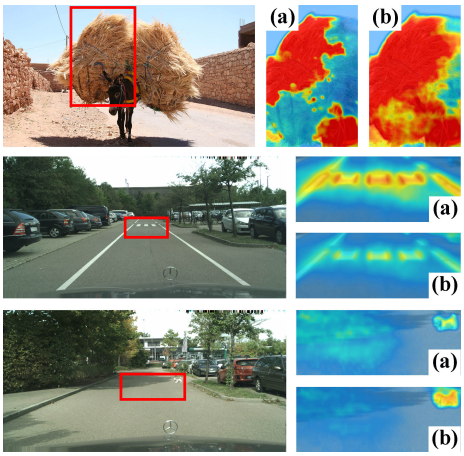

Fig. 7. **Qualitative comparisons on SML and BSL.** (a) shows anomaly scores from Standardized Max Logit (SML) while (b) shows anomaly scores from Blended Standardized Logits (BSL). The logits for (a) and (b) in each row are produced by the same model. Compared with SML, BSL can reduce false negative (row 1), false positive (row 2), and both (row 3).

Fig. 8. **Qualitative comparisons on different inference scales.** (a) indicates using only the original scale while (b) indicates using multiple scale inference. Other configurations are aligned. Multi-scale inference can mitigate the unexpected false positive existing in a specific inference scale (the false positive region of the sidewalk in row 1) and can better identify distant obstacles (row 2).

V. CONCLUSION

In this work, we propose an unsupervised road anomaly detection framework with language anchors from vision-language pretraining. We identify the importance of the pairwise binary logits and the negation anchor for representation transfer and present the Blended Standardized Logit as a novel strategy for uncertainty measurement. The effectiveness and robustness of this framework is manifested through the superior performance on diverse datasets for urban driving scenes. We believe that the proposed method demonstrates the potential competence of language-driven anomaly detection in various real-world contexts for autonomous agents.

REFERENCES

[1] M. Han, Z. Zhang, Z. Jiao, X. Xie, Y. Zhu, S.-C. Zhu, and H. Liu, "Scene reconstruction with functional objects for robot autonomy," *International Journal of Computer Vision*, vol. 130, no. 12, pp. 2940–2961, 2022.

[2] B. Jin, B. Tian, H. Zhao, and G. Zhou, "Language-guided semantic style transfer of 3d indoor scenes," in *Proceedings of the 1st Workshop on Photorealistic Image and Environment Synthesis for Multimedia Experiments*, 2022, pp. 11–17.

[3] X. Chen, H. Zhao, G. Zhou, and Y.-Q. Zhang, "Pq-transformer: Jointly parsing 3d objects and layouts from point clouds," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2519–2526, 2022.

[4] T. Zhang, X. Chen, Y. Wang, Y. Wang, and H. Zhao, "Mutr3d: A multi-camera tracking framework via 3d-to-2d queries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4537–4546.

[5] L. Yuan, X. Gao, Z. Zheng, M. Edmonds, Y. N. Wu, F. Rossano, H. Lu, Y. Zhu, and S.-C. Zhu, "In situ bidirectional human-robot value alignment," *Science Robotics*, 2022.

[6] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi, "Hoi4d: A 4d egocentric dataset for category-level human-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 013–21 022.

[7] X. Chen, T. Liu, H. Zhao, G. Zhou, and Y.-Q. Zhang, "Cerberus transformer: Joint semantic, affordance and attribute parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 649–19 658.

[8] B. Tian, L. Luo, H. Zhao, and G. Zhou, "Vibus: Data-efficient 3d scene parsing with viewpoint bottleneck and uncertainty-spectrum modeling," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 194, pp. 302–318, 2022.

[9] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S.-C. Zhu, "Diffusion-based generation, optimization, and planning in 3d scenes," *arXiv preprint arXiv:2301.06015*, 2023.

[10] P. Li, B. Tian, Y. Shi, X. Chen, H. Zhao, G. Zhou, and Y.-Q. Zhang, "Toist: Task oriented instance segmentation transformer with noun-pronoun distillation," in *Advances in Neural Information Processing Systems*.

[11] D. Bogdoll, M. Nitsche, and J. M. Zöllner, "Anomaly detection in autonomous driving: A survey," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4488–4499.

[12] B. Sun, J. Xing, H. Blum, R. Siegwart, and C. Cadena, "See yourself in others: Attending multiple tasks for own failure detection," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8409–8416.

[13] V. Sindhwani, H. Sidahmed, K. Choromanski, and B. Jones, "Unsupervised anomaly detection for self-flying delivery drones," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 186–192.

[14] D. J. Samuel and F. Cuzzolin, "Unsupervised anomaly detection for a smart autonomous robotic assistant surgeon (saras) using a deep residual autoencoder," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7256–7261, 2021.

[15] D. Park, Y. Hoshi, and C. C. Kemp, "A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1544–1551, 2018.

[16] H. Wang, Y. Sun, and M. Liu, "Self-supervised drivable area and road anomaly segmentation using rgb-d data for robotic wheelchairs," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4386–4393, 2019.

[17] L. Wellhausen, R. Ranftl, and M. Hutter, "Safe robot navigation via multi-modal anomaly detection," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1326–1333, 2020.

[18] I. Bozcan, J. Le Fevre, H. X. Pham, and E. Kayacan, "Gridnet: Image-agnostic conditional anomaly detection for indoor surveillance," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1638–1645, 2021.

[19] D. Mantegazza, A. Giusti, L. M. Gambardella, and J. Guzzi, "An outlier exposure approach to improve visual anomaly detection performance for mobile robots," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 354–11 361, 2022.

[20] L. Sun, K. Yang, X. Hu, W. Hu, and K. Wang, "Real-time fusion network for rgb-d semantic segmentation incorporating unexpected obstacle detection for road-driving images," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5558–5565, 2020.

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

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

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

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

[25] P. Bevandić, I. Krešo, M. Oršić, and S. Šegvić, "Simultaneous semantic segmentation and outlier detection in presence of domain shift," in *German conference on pattern recognition*. Springer, 2019, pp. 33–47.

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

[27] C. Creusot and A. Munawar, "Real-time small obstacle detection on highways using compressive rbm road reconstruction," in *2015 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2015, pp. 162–167.

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

[29] F. K. Gustafsson, M. Danelljan, and T. B. Schon, "Evaluating scalable bayesian deep learning methods for robust computer vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[30] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

[31] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[32] Y. Gal, J. Hron, and A. Kendall, "Concrete dropout," *Advances in neural information processing systems*, vol. 30, 2017.

[33] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," 2022, pp. 8759–8773.

[34] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[35] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," in *International Conference on Learning Representations*, 2020.

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

[37] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "Denseclip: Language-guided dense prediction with context-aware prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 082–18 091.

[38] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu, "Cris: Clip-driven referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 686–11 695.

[39] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *International Conference on Learning Representations*, 2021.

[40] M. Najibi, B. Singh, and L. S. Davis, "Autofocus: Efficient multi-scale inference," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9745–9755.

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

[42] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena, "Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

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

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

[45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[46] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.