2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

One-vs-All Semi-Automatic Labeling Tool for Semantic Segmentation in Autonomous Driving

Jing Gu[1], Guillermo Gallego[2], Amine Ben Arab[3]

***Abstract*— Semantic image segmentation plays a pivotal role in creating High-Definition (HD) maps for autonomous driving, where every pixel in an image is assigned a label from a specific semantic class. However, obtaining dense pixel-level annotations for model training is a laborious and expensive process. Active learning holds promise as a method to reduce the human annotation effort needed for semantic segmentation. However, existing active learning methods often perform well in the majority classes but struggle with the minority classes, negatively impacting segmentation performance. To tackle this challenge, we propose a novel One-vs-All (OVA) active learning framework, known as *OVAAL*. This paper explains how *OVAAL* can shift more attention towards the minority classes and thoroughly analyzes its contributions to performance enhancement. Additionally, we introduce an OVA-based semi-supervised learning method as the final training phase, referred to as *OVAAL+*. Our results demonstrate that both *OVAAL* and *OVAAL+* lead to significant improvements, with mean Intersection over Union (mIoU) gains of 4.55% and 6.38%, respectively, compared to the state-of-the-art active learning method *Pixelpick* on the Cityscapes semantic segmentation benchmark. These improvements are achieved while maintaining an economical annotation budget of 1.44% of the training data. We foresee further research exploring the potential of OVA-based active selection to address challenges in cold start scenarios and resource-constrained training environments.**

I. Introduction

Cameras are fundamental components in perception systems for autonomous driving and rank among the most prevalent and indispensable sensors in autonomous vehicles [1]. In this scenario, scene understanding is paramount, and image semantic segmentation is widely adopted as a key task [2].

One of the main challenges in training an accurate semantic image segmentation model is the acquisition of sufficient data. The conventional method of manual pixel-level annotation, while effective, is both time-consuming and costly. For instance, in the Cityscapes segmentation dataset [3], human annotators spent an average of 1.5 hours to label each image. Similarly, in the Mapillary Vistas dataset, the annotation process took approximately 94 minutes per image [4]. The cost of manual annotation can become prohibitive, especially for large datasets. With the cost of labeling one segmentation polygon estimated at 0.84 dollars based on Amazon's data labeling pricing [5], creating a dataset with an average of 30 polygons per image for a dataset with 10000 images could result in a staggering total cost of 252000 dollars. This financial burden highlights the urgent need for more cost-effective solutions.

Fig. 1. Visualization of queried pixels to annotate (1.44%) on Cityscapes dataset using different One-vs-All (OVA) models. The OVA pixel-based active selection strategy can bring more diversity. Each OVA model only selects the samples related to its own classes (represented with different colors), forcing the selection to focus more on the minority classes. Best viewed when zoomed in.

In response to this challenge, researchers have proposed various solutions to reduce human annotation efforts and enhance the efficiency of semantic segmentation [6]–[15]. Recently, a promising approach has emerged: active learning [8], [9], [12], [15]. Active learning strategically selects the most informative samples for model training. It has demonstrated the ability to significantly reduce human annotation efforts while achieving performance on par with fully supervised learning, even with a reduced subset of the training data [16].

However, traditional active learning methods for semantic segmentation often excel in the majority classes while they struggle with the minority classes, resulting in compromised segmentation performance. This issue can be attributed to a lack of selection diversity. For example, *Pixelpick* [13] suggests that using sparse pixel labels can introduce more spatial diversity into the selection process. *Superpixels* [17] estimates the class distribution within individual superpixel regions and, based on this distribution, reduces the selection of majority classes while prioritizing the query of rare samples. In a similar way, the domain adaptation work *RIPU* [14] introduces the concept of region impurity to improve the semantic diversity of the selected samples.

While region-based selection methods like *Superpixels* and *RIPU* provide enhanced semantic diversity, they often lack the spatial diversity present in pixel-based selection. This limitation can lead to increased annotation costs. On the other hand, pixel-based selection methods, such as *Pixelpick*, may perform poorly in the minority categories due to the limited semantic diversity in their selection process.

Building on these observations, we propose a novel One-

[1]Jing Gu (Corresponding Author) is with Expleo Germany GmbH, Salufer 8, 10587 Berlin, Germany jing.gu@expleogroup.com
[2]Guillermo Gallego is with the ECDF and TU Berlin, Berlin, Germany.
[3]Amine Ben Arab is with Expleo Germany GmbH, Wilhelm-Wagenfeld-Str.1-3, 80807 München, Germany. amine.ben-arab@expleogroup.de

vs-All (OVA) pixel-based active learning framework, denoted as *OVAAL*. This framework decomposes multi-class classification into several binary classifications, emphasizing a more equitable treatment of different classes. We argue that OVA-based active selection compels the model to pay greater attention to the minority classes. Additionally, we observe that the prediction of OVA is binary, which makes using the unreliable (high-uncertainty) prediction easier compared to the standard multi-class classification [11]. We further propose a semi-supervised learning-based method, termed *OVAAL+*, building upon our OVA active learning framework. Our experiments demonstrate that OVA-based active learning enhances the accuracy of minority class predictions while maintaining an affordable annotation budget. Furthermore, our OVA-based semi-supervised learning proves beneficial for overall segmentation performance.

In summary, our contributions can be outlined as follows:

- We introduce a novel One-vs-All based Active Learning framework, *OVAAL*, which places increased emphasis on the minority classes, thereby improving segmentation performance.
- We propose a One-vs-All based semi-supervised learning method as an extension for our OVA active learning framework, *OVAAL+*, which further enhances segmentation performance.
- We conduct a thorough experimental analysis of the contributions of our OVA-based sample selector, classifier, and semi-supervised learning to the achieved performance improvements.

II. Related Work

A. Active learning for semantic segmentation

The fundamental hypothesis of active learning is that if the learning algorithm can choose the data from which it learns, it will perform better with less training data [16]. Active learning is a selection problem that aims to identify and utilize the most informative samples for training. Active selection methods can be broadly categorized into two types: uncertainty-based and representation-based [15].

VAAL [9], *FeatureMixing* [18], *CoreSet* [12] are typical representation-based active learning methods. However, these methods are usually evaluated on image classification benchmarks [15]. Uncertainty-based methods have proven to be effective in semantic segmentation, as demonstrated by approaches such as *EquAL* [8], *DEAL* [15], *Superpixels* [17], *RIPU* [14] and *Pixelpick* [13]. However, uncertainty alone may not capture the entire diversity of the data, which is crucial for practical training [19]. Thus, uncertainty is often combined with other techniques. For instance, *RIPU* leverages impurity in square regions with prediction uncertainty to capture diverse samples, while *DEAL* utilizes the difficulty associated with uncertainty to prioritize the selection of samples that challenge the model's predictions.

Following the selection framework of *Pixelpick* [20] we introduce OVA into the active selection process, which leads to different uncertainty estimation, training objectives, and inference. We discuss these differences in Sec. III.

B. One-vs-All (OVA) Classification

One-Versus-All (OVA) classification is a widely adopted method for addressing multi-class classification tasks [21]. OVA decomposes a multi-class problem with m classes into m binary sub-problems. As the number of categories increases, the number of base classifiers required for OVA increases linearly.

Previous research [22]–[26] has demonstrated that the One-Versus-All (OVA) approach is an effective decomposition method for solving multiclass classification tasks, maintaining performance comparable to standard multi-class classification. However, when applying OVA for pixel-level classification in semantic segmentation, even in the absence of class imbalance among the classes, an imbalance of positive and negative samples arises due to the larger number of other classes compared to the target class [21]. This class imbalance poses a significant challenge.

To address the class imbalance problem introduced by the decomposition, Zhang et al. [27] proposed a One-Versus-One (OVO) based ensemble learning approach with preprocessing techniques. However, as the number of categories increases, the number of base classifiers required for OVO grows quadratically. GAO et al. [21] introduced a differential partition sampling method, dividing all samples into four categories: safe examples, borderline examples, rare examples, and outliers, based on neighborhood information. During sampling iterations, they undersampled or oversampled the majority and minority classes in each binary training dataset to balance the number of positive and negative samples.

Our work shares similarities with the idea presented in [21], as we also address class imbalance by upsampling and oversampling pixel samples in a single image to construct a binary dataset. However, our approach differs in that our sampling process is driven by the model itself, eliminating the need for manual definitions of different sample categories. Additionally, when dealing with the class imbalance problem inherent in OVA, our active learning-based sampling method requires fewer computational resources than OVO.

C. Semi-supervised learning for semantic segmentation

Semi-supervised learning represents a family of algorithms trained on labeled samples that seek to learn from the unlabeled data. The labeled data and unlabeled data form the full dataset. It assumes that both types of data are sampled from the same or similar distributions [28]. Semi-supervised learning typically employs consistency regularization and Pseudo-Labeling [28].

Pseudo-labeling involves selecting reliable pseudo-labels using a confidence threshold. Wang et al. [11] separate reliable and unreliable pixels based on entropy. *S4L* [28] uses predictions with high confidence as pseudo-labels to refine the models. *Cutout* [29] is a simple regularization technique for convolutional neural networks that involves removing contiguous sections of input images. *FixMatch* [30] follows a similar approach, generating one weakly-augmented image and one strongly-augmented image and training the model

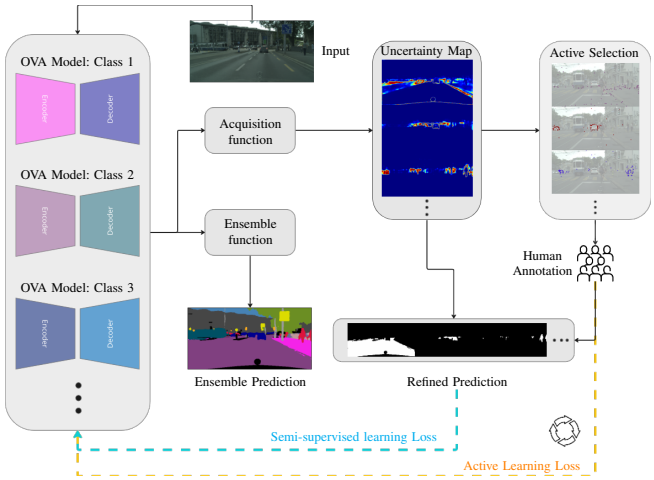

Fig. 2. Overview of our One-versus-All active learning (OVAAL) framework for semantic image segmentation. At each round of active learning, we utilize each OVA model's prediction uncertainty to select a batch of pixels and query annotation from a human. Finally, the network is retrained using all labeled data. The semi-supervised learning step combines the human annotation and prediction uncertainty to refine the pseudo labels, which are used to retrain the network.

to match the predictions on the strongly-augmented version with those on the weakly-augmented one.

Our work goes in the direction of using Pseudo-Labeling. Different from the multi-class Pseudo-Labeling. For each OVA model, the prediction is binary, and we can use uncertainty to invert the prediction.

III. PROPOSED APPROACH: OVAAL

Our objective is to train a highly accurate model for semantic segmentation using *pool-based active learning* [16]. In this approach, we initially train a model on a small set of labeled data and then request labels from a large pool of unlabeled data. Our work draws inspiration from *Pixelpick* [13], which operates under the assumption that even a few paired pixel samples from images and their corresponding labels can effectively train a high-quality semantic segmentation network. However, unlike *Pixelpick*, our approach employs One-Vs-All (OVA) classification to replace the standard multi-class classifier, leading to a novel retraining process for the segmentation model and a unique sampling strategy. Additionally, we leverage the uncertainty estimation provided by the OVA models to refine the pseudo-labeling process.

As illustrated in Fig. 2, our active learning branch utilizes the input images for the OVA models. The predictions generated by the OVA models offer pixel-wise uncertainty estimates for each input image. High-uncertainty pixel samples are then selected for annotation by humans to obtain ground truth labels. These annotations are used to retrain the OVA models. In the semi-supervised learning branch, the OVA prediction is binary, a characteristic that facilitates the use of uncertainty for prediction refinement. We can invert the binary prediction when we assume that highly uncertain predictions are likely to be incorrect. For example, if a pixel's prediction is foreground and it exhibits high uncertainty, we can refine the prediction to classify it as background. The refined predictions, along with previous human annotations, serve as pseudo labels for the retraining of the OVA model.

Formally, in our OVA active learning framework, $I \subset \mathbb{R}^{H\times W\times 3}$ represents the space of color images. $Y_c \subset \mathbb{R}^{H\times W}$, where $c \subset C$ and $(Y_c(u,v) = 0 \text{ or } 1)$ denotes binary pixel-wise labels belonging to a specific class at pixel location (u, v). Here, $n(C)$ represents the total number of categories, and c represents a specific category. The objective is to learn a function $\Phi_c(\theta) : I \rightarrow Y_c$. $\Phi_c(\theta)$ represents an OVA ConvNet with parameters θ that maps an input image to the label space Y_c. For categories c, D_U represents the unlabeled database, and D_L denotes the labeled database. We assume a B pixel samples annotation budget, which is consumed in K selection rounds or iterations.

Active Learning: At the k^{th} selection round, in the first step, unlabeled image data I creates the unlabeled database D_U^k. Moving on to the second step, a neural network $\Phi_c(\theta)^{k-1}$ is trained with the last round of annotated labels, which maps the input image data I to $P_c \subset \mathbb{R}^{H\times W\times 2}$, where P_c represents the binary probability of the class c. The third step involves sampling H high uncertainty pixel samples using the acquisition function A, denoted as $A(D_U, \Phi_c(\theta), H)$. This function utilizes the neural network $\Phi_c(\theta)^{k-1}$ and the unlabeled database D_U to generate candidate pixel coordinates h_c^k. Subsequently, in the fourth step, the pixel coordinates h_c^k are provided to human annotators to obtain the ground truth for these selected pixels, denoted as d_l^k. The fifth step involves using the annotations to construct an annotated database D_L. Finally, in the sixth step, the neural network $\Phi_c(\theta)^{k-1}$ leverages the labeled data D_L to optimize its parameters and generate the new model $\Phi_c(\theta)^k$. Steps 2 to 6 are repeated iteratively until the annotation budget is exhausted. In the initial stages, to bootstrap the system when the neural network is not trained, the acquisition function randomly selects pixels from an image as candidates.

Semi-Supervised learning: After consuming the entire annotation budget, the acquisition function A selects the top L low confidence (high uncertain) samples l_c and assigns labels to these selected samples using the inverse predictions of the model $\Phi_c(\theta)^K$. We calculate the active learning loss on the Labeled database D_L and semi-supervised learning loss on the pseudo-labels l_c. The model uses these two losses to update its neural network weights and generate the new model $\Phi_c(\theta)_S^k$.

The whole acquisition process is defined by the equation:

$$\mathcal{S} = \begin{cases} h_c^k = A(D_U, \Phi_c^{k-1}(\theta), H) & \text{if active selection,} \\ l_c = A(D_U, \Phi_c^{k-1}(\theta), L) & \text{if post processing.} \end{cases} \quad (1)$$

Acquisition function: Since the focus of our work is not on the design of another criterion, but rather on the effectiveness of OVA as the basic classifier for active learning, we consider the existing approach based on the framework of uncertainty sampling [16], with three commonly used uncertainty estimation methods: *Least Confidence*, *Margin*

Sampling and *Entropy*. Because of the OVA approach, all three uncertainty estimation methods reduce to the same one in binary classification scenarios. Hence, we use *Entropy* to measure the uncertainty of each pixel in one image. The *Entropy* method is expressed as follows:

$$x_E = -\sum_i P_\theta(y_i|x) \log P_\theta(y_i|x) \tag{2}$$

$$A(D_U, \Phi_c^{k-1}(\theta), N) = \{x_E^1, x_E^2, \cdots, x_E^N\},$$
$$\text{where } \{x_E^1 \geq x_E^2 \geq \cdots \geq x_E^k, x_E \in D_U\}.$$

Active learning Loss function: We use the cross-entropy loss to optimize the weights in the neural network. At the k^{th} selection round, the expanded database D_L^k and the model $\Phi_c(\theta)^{k-1}$ from the previous round ($(k-1)^{\text{th}}$) are available. Here, $\|D_L^k\|$ represents the number of samples in the expanded database. The optimization of the parameters θ_c involves minimizing the cross-entropy loss on the current expanded database D_L^k for c. Unlike the standard cross-entropy loss, we only calculate the loss at the labeled pixel coordinates. The mathematical expression for this loss function is as follows:

$$L_{\text{active}} = -\frac{1}{\|D_L^k\|} \sum_{(u,y_u)\in D_L^k} \big(y_u(c) \log(\hat{y}_u(c)) + (1-y_u(c)) \log(1-\hat{y}_u(c))\big),$$

where $y_u(c)$ and $\hat{y}_u(c)$ denote the annotated label and corresponding model prediction at pixel coordinate u, specifically for the category c, represented in binary format.

Semi-supervised learning Loss function: The top L high uncertain samples l_c are selected by the A Acquisition function. The prediction's probability of the input data is $P_c(x)$, and the highest probability channel indicates whether the prediction is either background or foreground. We use the selected high-uncertainty samples l_c to refine the prediction.

$$R_c(x) = \begin{cases} \arg\max(P_c^i(x)),\ i \in [0,1] & \text{if } x \notin l_c \\ \arg\min(P_c^i(x)),\ i \in [0,1] & \text{if } x \in l_c \end{cases} \tag{3}$$

To calculate the semi-supervised loss between the refined prediction and the original prediction, we employ the cross-entropy loss function defined as follows:

$$L_{\text{semi}} = -\frac{1}{N} \sum_{i=1}^{N} \big(\hat{y}_i(c) \log(p_i(c)) + (1-\hat{y}_i(c)) \log(1-p_i(c))\big),$$

where N represents the number of pixels in an image.

To fully utilize the annotated data, we combine the semi-supervised loss with the loss of annotated data, resulting in the following overall loss function:

$$L = L_{\text{active}} + \alpha L_{\text{semi}}. \tag{4}$$

The hyper-parameter α controls the weight of the semi-supervised learning component.

Ensemble method: With OVA, we can decouple a multi-class classification problem into several binary classification problems. To predict the multi-class semantic segmentation, we need to ensemble the results of different binary classifiers. We follow the Winner-Takes-All (WTA) [22] ensemble method, which can be defined as follows:

$$P(u,v) = \arg\max_c P_c(u,v) \tag{5}$$

The WTA strategy assigns the samples to the class with the highest membership score. $P_c(u,v) \in [0,1]$ represent the prediction probability of model $\Phi_c(\theta)$.

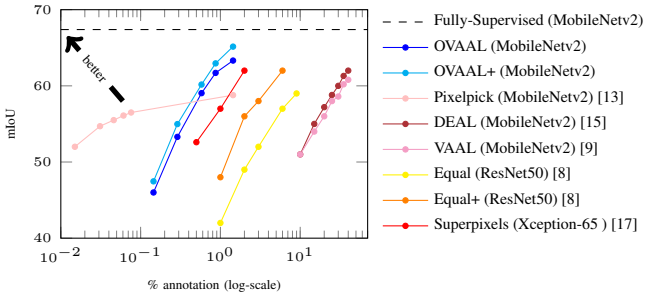

Fig. 3. Comparison with the state of the art using the Cityscapes dataset.

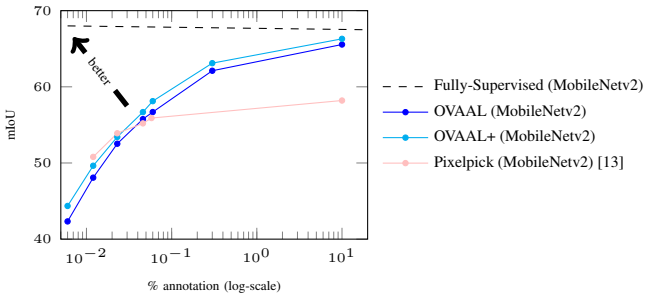

Fig. 4. Comparison with *Pixelpick* using the CamVid dataset.

IV. EXPERIMENTS

Datasets. Cityscapes is a dataset for semantic urban scene understanding [3]. It collects data from various cities in Germany. The original resolution of the image data (1024×2048 px) provides more information for training. The dataset includes 19 annotated classes for semantic segmentation. The dataset is divided into 2975 training samples and 500 validation samples. Camvid [31] is a dataset for the urban semantic segmentation scenario. Data from Camvid are collected from the perspective of a driving automobile. The original resolution of Camvid is (720×960 px). Camvid has 367 samples for training and 233 samples for evaluation.

Training settings. We train all methods with DeepLabv3+ models [32] on Cityscapes [3] and Camvid [31] with the encoder backbone MobileNetv2 [33]. The training batch size is 16. We set the learning rate to 5×10^{-4} using the Adam optimizer [34]. We use Adam's default settings from *Pixelpick*, where the exponential decay rates for the moment estimates β_1 and β_2 are set to 0.9 and 0.999, respectively. Additionally, we set the weight decay as 2×10^{-4}. For the semi-supervised learning based step, we set the hyper-parameter $\alpha = 1$. Data augmentation is applied to improve model generalization and robustness.

TABLE I

PER-CLASS IOU OF AND MIOU [%] ON CITYSCAPES ORIGINAL VALIDATION. THE BEST VALUES PER COLUMN ARE HIGHLIGHTED IN **BOLD**.

Classes Label frequency (%)	road 24.299	sidewalk 3.607	building 12.083	wall 0.406	fence 0.582	pole 0.772	tr. light 0.204	tr. sign 0.360	veget. 8.916.	terrain 0.735	sky 1.805	person 0.929	rider 0.128	car 4.758	truck 0.179	bus 0.169	train 0.157	m/cycle 0.069	bicycle 0.297	**mIoU**
VAAL (MobileNetv2 40%) [9]	96.22	73.27	86.95	**47.27**	43.92	37.40	36.88	54.90	87.10	54.48	91.63	63.44	38.92	87.92	50.15	63.70	52.36	35.99	54.97	60.92
DEAL (MobileNetv2 40%) [15]	95.89	71.69	87.09	45.61	**44.94**	38.29	36.51	55.47	87.53	**56.90**	91.78	64.25	39.77	88.11	56.87	64.46	50.39	38.92	56.69	61.64
Pixelpick (MobileNetv2 **1.44**%) [13]	94.07	68.49	86.57	37.15	40.67	42.64	40.90	55.32	87.39	48.54	88.42	68.00	42.00	89.49	38.61	55.20	37.94	32.71	62.67	58.78
OVAAL (MobileNetv2 **1.44**%) (Ours)	96.73	**76.85**	88.00	36.09	35.14	**47.87**	47.80	**61.47**	88.82	44.73	**92.97**	**71.03**	45.69	**92.21**	50.41	68.19	46.49	31.63	**66.01**	63.33
OVAAL+ (MobileNetv2 **1.44**%) (Ours)	**96.77**	75.83	**88.66**	38.59	41.16	47.28	**47.99**	60.37	**89.80**	49.03	92.80	70.83	**47.77**	91.51	**58.53**	**73.53**	**59.92**	**41.76**	65.78	**65.16**

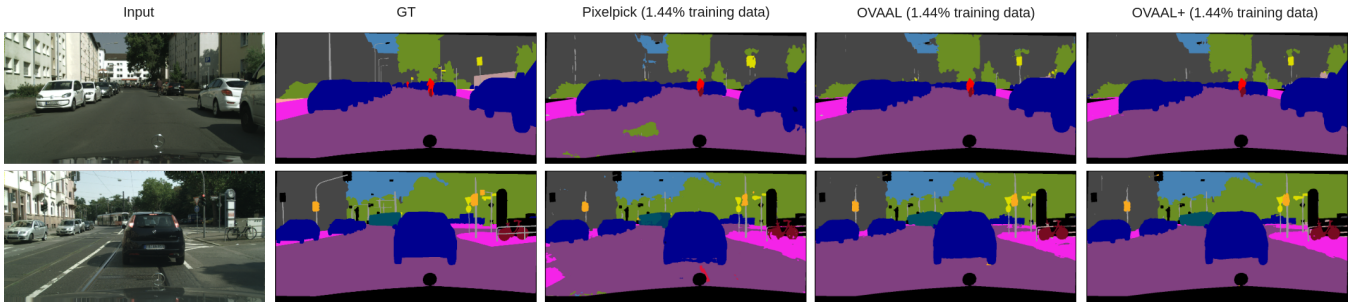

Fig. 5. Qualitative segmentation results on the Cityscapes dataset.

A. Main Results

We conduct a comparative analysis of our approach against several state-of-the-art (SOTA) active learning models for semantic segmentation, including *VAAL* [9], *DEAL* [15], and *Pixelpick* [13]. Table I presents a per-class Intersection over Union (IoU) comparison between our model and these SOTA approaches using the Cityscapes dataset [3]. Figure 3 and Fig. 4 illustrate the segmentation performance of different active learning methods under varying annotation budgets on the Cityscapes and CamVid datasets. Furthermore, Fig. 5 provides qualitative results on the Cityscapes dataset.

We use *OVAAL* to denote our method without the semi-supervised learning, while *OVAAL+* indicates that we employ refined predictions as pseudo labels for model retraining.

OVAAL vs. Standard Multi-class Based Pixel Selection: Our *OVAAL* follows the pixel-based selection framework of *Pixelpick*. Compared to *Pixelpick*, our *OVAAL* method utilizes the same annotation budget (1.44% of the training data) but exhibits a 4.55% improvement (see Tab. I) in semantic segmentation performance (mIoU). Notably, in Tab. I, our *OVAAL* excels in the minority classes (classes with less than 0.5% of the pixels in the training data). For instance, the class "traffic light" comprises only 0.2% of the training data pixels, yet our *OVAAL* demonstrates a 6.97% improvement in accuracy for this class. Similar improvements are observed for other minority classes such as "rider" (0.12% px), "truck" (0.179%), "bus" (0.169%), and "train" (0.157%). These results support our assumption that *OVAAL* allocates equal attention to minority classes, resulting in substantial accuracy improvements. To substantiate this claim, we visually analyzed the selected points from various OVA models in Fig. 1. The figure illustrates that OVA's active selection is specific to its categories, focusing on selecting the most informative samples related to each class.

OVAAL vs. Standard Multi-class Based Image Selection: *VAAL* and *DEAL* employ image-based selection frameworks. In comparison, our *OVAAL* is trained with only 1.44% of the training data, while *VAAL* and *DEAL* use 40% of the training data. Our *OVAAL* outperforms *VAAL* by 2.41% and *DEAL* by 1.69% in terms of mIoU (see Tab. I). The pixel-based selection framework allows the model to utilize more image data with fewer annotations compared to image-based selection. For example, with an annotation budget of 1900 pixels per image, an image size of 256×512 px, and a training dataset of 2975 images, our method uses only 1.44% of the data, equivalent to approximately 43 whole images in the image-based selection method. The pixel-based selection framework inherently introduces more diversity into the selection process, inheriting this advantage from [13]. Moreover, even with a smaller annotation budget, our *OVAAL* maintains superior prediction accuracy, particularly in the minority classes.

TABLE II

MIOU [%] ON CITYSCAPES ORIGINAL VALIDATION.

Methods	mIoU
(a) Pixelpick + Pixelpick selection	58.50
(b) Pixelpick + OVAAL selection (Ours)	61.04
(c) OVAAL (Ours) + Pixelpick selection	61.49
(d) OVAAL (Ours) + OVAAL selection (Ours)	63.33
(e) OVAAL (Ours) + OVAAL selection (Ours) + Semi-supervised learning	65.15

B. Ablation Studies

To gain a deeper understanding of the individual components of our *OVAAL* framework, we conducted ablation studies on the Cityscapes dataset. Table II summarizes the results, with all methods using only 1.44% of the training data.

Effect of OVA Selection: In Table II, we compare the performance of models (a) and (b), both using a standard multi-class-based classifier. To ensure a fair comparison, both models are trained for 150 epochs from scratch. The key difference between them is the selected pixels. Model (b) utilizes pixel selection from *OVAAL*, resulting in a significant improvement compared to model (a). This observation underscores the value of our OVA-based pixel selection in providing valuable information to the model. Furthermore, in

TABLE III
SPATIAL DIVERSITY AND SEMANTIC DIVERSITY OF SELECTION.

Methods	Spatial diversity ↑	Semantic diversity ↑
(a) Pixelpick	138.46	3.3917
(b) OVAAL (Ours)	**141.63**	**3.4397**

model (c), we employ pixel selection from *Pixelpick* to train our OVA-based model, and we notice a performance decrease compared to model (d), which directly utilizes OVA-based selection. These results reaffirm that information from OVA-based selection is advantageous for training both standard multi-class and OVA-based models.

Effect of OVA Classifier: Models (a) and (c) in Tab. II, as well as models (b) and (d), share the same selection for training. In both cases, introducing the OVA classifier yields an improvement of more than 2% mIoU on the Cityscapes dataset. This demonstrates that the OVA classifier exhibits superior learning capabilities.

Effect of Semi-Supervised Learning: As discussed in Section III, we contend that high uncertainty in predictions indicates potential unreliability. Leveraging the OVA binary predictions, we simply invert the unreliable binary predictions. Model (e) exhibits a 1.82% improvement in mIoU compared to model (d), highlighting the effectiveness of our OVA-based semi-supervised learning strategy in enhancing the final segmentation performance.

C. Diversity of Selection

We visualize the selection of three OVA models in Fig. 1. We find that the selection of model "train" focuses on the pixels which are highly related to the object "train". This selection forces the model to put the same effort into majority and minority categories, which leads to better performance in the rare categories. To further compare the diversity of selection, we calculate the entropy of the selection's class distribution as the semantic diversity. We calculate the standard deviation of the selected pixels' coordinates as spatial diversity. As shown in Tab. III, our method brings more spatial and semantic diversity compared to *Pixelpick*.

D. Drawbacks of Our OVAAL

Difficulty in Cold Start: Our *OVAAL* method faces challenges during the cold start when the annotation budget is limited. In the initial selection round, we employ random selection to train our model. However, for minority classes, it can be challenging to obtain enough training data for initialization. In standard multi-class selection, even if it doesn't select a sufficient number of minority class samples, the probabilities of other classes can provide some information about unseen or minority class samples. In contrast, our OVA-based selection treats each category separately and equally, promoting selection diversity when the annotation budget is sufficient (e.g., 1.44% of the training data). However, as illustrated in Fig. 3, when the annotation budget is extremely limited (e.g., 0.076% of the training data), our OVA-based selection may struggle to select an adequate number of training samples for minority classes, resulting in performance lag compared to the standard multi-class selection method *Pixelpick*. A similar trend is observed within the Camvid dataset analysis (e.g., Fig. 4). *Pixelpick* shows better performance when the proportion of annotated pixels is below 0.023% of the training samples. Conversely, an increase in the proportion of annotated pixels to beyond 0.06% marks a significant enhancement in segmentation performance through our OVAAL approach.

Increased Training Resource Requirements: Our *OVAAL* method demands more training resources compared to the standard multi-class selection method. OVA decomposes a multi-class problem with m classes into m binary sub-problems. As the number of categories increases, the number of base classifiers required for OVA grows linearly, imposing a greater computational burden.

Limitations of Ensemble Strategy: As outlined in Sec. III, we employ the Winner-Takes-All (WTA) strategy to ensemble the predictions from multiple binary models into a multi-class prediction. However, this approach treats each category independently, neglecting potential mutual information between different categories. For instance, if the predicted probability for the "Road" category is 0.8, and the predicted probability for the "Sky" category is 0.79, the WTA strategy assigns a final probability of 0.8. This simplification can lead to inaccuracies in the final predictions. The exploration of a better ensemble strategy that accounts for the interdependencies between categories represents a valuable topic for future research.

V. CONCLUSIONS

This work has introduced an innovative One-vs-All (OVA) based active learning algorithm tailored for efficient semantic segmentation model training using sparsely annotated pixels. Our study has revealed that OVA-based active selection enhances selection diversity, and the OVA classifier demonstrates improved performance when coupled with active learning. Furthermore, we have conducted comprehensive comparisons with existing state-of-the-art active learning approaches on the Cityscapes dataset, demonstrating the superiority of our OVA-based approach while requiring an affordable number of annotations.

To the best of our knowledge, our work represents the first exploration of OVA combined with active learning in the context of semantic segmentation. We anticipate that our research will inspire further investigations into the promising application of OVA-based active selection. Our contributions aim to address the challenges associated with annotation efficiency in semantic segmentation, contributing to the advancement of autonomous driving systems and other computer vision applications.

ACKNOWLEDGMENT

The research leading to these results is funded by the Bavarian Ministry of Economic Affairs, Energy and Technology and by Expleo Germany GmbH within the project SAFE2P (DIK-350). The authors would like to thank Alireza Ferdowsizadeh Naeeni for his support.

REFERENCES

[1] C. Gao, G. Wang, W. Shi, Z. Wang, and Y. Chen, "Autonomous driving security: State of the art and challenges," *IEEE Internet of Things Journal*, vol. 9, no. 10, pp. 7572–7595, 2021.

[2] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58443–58469, 2020.

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 3213–3223, 2016.

[4] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Int. Conf. Comput. Vis. (ICCV)*, pp. 4990–4999, 2017.

[5] "Amazon sagemaker data labeling pricing." https://aws.amazon.com/sagemaker/data-labeling/pricing/?nc=sn&loc=3. Accessed: 2023-07-29.

[6] J. H. Cho, U. Mall, K. Bala, and B. Hariharan, "Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 16794–16804, 2021.

[7] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Int. Conf. Comput. Vis. (ICCV)*, pp. 1635–1643, 2015.

[8] S. A. Golestaneh and K. M. Kitani, "Importance of self-consistency in active learning for semantic segmentation," *arXiv preprint arXiv:2008.01860*, 2020.

[9] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *Int. Conf. Comput. Vis. (ICCV)*, pp. 9865–9874, 2019.

[10] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 3159–3167, 2016.

[11] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 4248–4257, 2022.

[12] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," *arXiv preprint arXiv:1708.00489*, 2017.

[13] G. Shin, W. Xie, and S. Albanie, "All you need are a few pixels: semantic segmentation with pixelpick," in *Int. Conf. Comput. Vis. (ICCV)*, pp. 1687–1697, 2021.

[14] B. Xie, L. Yuan, S. Li, C. H. Liu, and X. Cheng, "Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 8068–8078, 2022.

[15] S. Xie, Z. Feng, Y. Chen, S. Sun, C. Ma, and M. Song, "Deal: Difficulty-aware active learning for semantic segmentation," in *Asian Conf. Comput. Vis. (ACCV)*, 2020.

[16] B. Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[17] L. Cai, X. Xu, J. H. Liew, and C. S. Foo, "Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 10988–10997, 2021.

[18] A. Parvaneh, E. Abbasnejad, D. Teney, G. R. Haffari, A. Van Den Hengel, and J. Q. Shi, "Active learning by feature mixing," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 12237–12246, 2022.

[19] Y. Siddiqui, J. Valentin, and M. Nießner, "Viewal: Active learning with viewpoint entropy for semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 9433–9443, 2020.

[20] "Pixlepick implementation." https://github.com/NoelShin/PixelPick. Accessed: 2023-07-29.

[21] X. Gao, Y. He, M. Zhang, X. Diao, X. Jing, B. Ren, and W. Ji, "A multiclass classification using one-versus-all approach with the differential partition sampling ensemble," *Engineering Applications of Artificial Intelligence*, vol. 97, p. 104034, 2021.

[22] G. Franchi, A. Bursuc, E. Aldea, S. Dubuisson, and I. Bloch, "One versus all for deep neural network for uncertainty (ovnni) quantification," *IEEE Access*, vol. 10, pp. 7300–7312, 2021.

[23] B. Krawczyk, M. Galar, M. Woźniak, H. Bustince, and F. Herrera, "Dynamic ensemble selection for multi-class classification with one-class classifiers," *Pattern Recog.*, vol. 83, pp. 34–51, 2018.

[24] S. Padhy, Z. Nado, J. Ren, J. Liu, J. Snoek, and B. Lakshminarayanan, "Revisiting one-vs-all classifiers for predictive uncertainty and out-of-distribution detection in neural networks," *arXiv preprint arXiv:2007.05134*, 2020.

[25] K. Polat and K. O. Koc, "Detection of skin diseases from dermoscopy image using the combination of convolutional neural network and one-versus-all," *J. Artificial Intell. and Syst.*, vol. 2, no. 1, pp. 80–97, 2020.

[26] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *J. Mach. Learning Research (JMLR)*, vol. 5, pp. 101–141, 2004.

[27] Z. Zhang, B. Krawczyk, S. Garcia, A. Rosales-Pérez, and F. Herrera, "Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data," *Knowledge-Based Systems*, vol. 106, pp. 251–263, 2016.

[28] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *Int. Conf. Comput. Vis. (ICCV)*, pp. 1476–1485, 2019.

[29] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[30] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 596–608, 2020.

[31] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Eur. Conf. Comput. Vis. (ECCV)*, pp. 44–57, Springer, 2008.

[32] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Eur. Conf. Comput. Vis. (ECCV)*, pp. 801–818, 2018.

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 4510–4520, 2018.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.