

SegXAL: Explainable Active Learning for Semantic Segmentation in Driving Scene Scenarios

Sriram Mandalika¹ and Athira Nambiar¹

Department of Computational Intelligence,
 Faculty of Engineering and Technology,
 SRM Institute of Science and Technology
 Kattankulathur, Tamil Nadu, 603203, India
 mc9991@srmist.edu.in, athiram@srmist.edu.in

Abstract. Most of the sophisticated AI models utilize huge amounts of annotated data and heavy training to achieve high-end performance. However, there are certain challenges that hinders the deployment of AI models “in-the-wild” scenarios i.e. inefficient use of unlabeled data, lack of incorporation of human expertise and lack of interpretation of the results. To mitigate these challenges, we propose a novel Explainable Active Learning (XAL) model *viz.* ‘**XAL-based semantic segmentation model “SegXAL”**’, that can (i) effectively utilize the unlabeled data, (ii) facilitate the “Human-in-the-loop” paradigm and (iii) augment the model decisions in an interpretable way. In particular, we investigate the application of the SegXAL model for semantic segmentation in driving scene scenarios. The SegXAL model proposes the image regions that require labelling assistance from Oracle by dint of explainable AI (XAI) and uncertainty measures in a weakly-supervised manner. Specifically, we propose a novel Proximity-aware Explainable-AI (PAE) module and Entropy-based Uncertainty (EBU) module to get an Explainable Error Mask, which enables the machine teachers/human experts to provide intuitive reasoning behind the results and to solicit feedback to the AI system, via an active learning strategy. Such a mechanism bridges the semantic gap between man and machine through collaborative intelligence, where humans and AI actively enhance each other’s complementary strengths. A novel high-confidence sample selection technique based on the DICE similarity coefficient is also presented within the SegXAL framework. Extensive quantitative and qualitative analyses are carried out in the benchmarking Cityscape dataset. Results show the outperformance of our proposed SegXAL against other state-of-the-art models.

Keywords: Active learning · Explainable AI · Semantic segmentation.

1 Introduction

Over the past decade, the world has witnessed an unprecedented technological revolution with the help of Artificial Intelligence (AI) towards accelerating

automation, improving decision-making processes, and extracting insights from vast datasets. Despite these advancements, deep learning models commonly encounter substantial challenges while deploying in real-world or “in-the-wild” settings, such as limitation of well-annotated data, contextual & prior information and interpretability of the results [1].

Annotation of new data points is an expensive and laborious task, yet crucial for enriching training datasets with valuable information. In tasks like image semantic segmentation, manually labelling each pixel with its class label is arduous. Supervised algorithms provide efficient solutions for this task, whereas in unsupervised scenarios, automatic labelling poses a significant challenge for machines. Furthermore, *integrating prior and contextual information* can significantly enhance AI model performance, especially in high-risk scenarios e.g. medical and defence. Domain experts can contribute valuable knowledge to AI systems in such situations, enabling a “Human-in-the-loop” paradigm for more rational analysis and informative results. However, most existing AI systems lack mechanisms to incorporate *additional human-collected information or domain expertise*. In real-world scenarios, the inverse situation also exists, wherein the operators often have to rely on visual inspection to make decisions due to the *lack of explainability in machine decisions*. Despite the advancements in deep neural networks, the integration of AI tools in various fields is hindered by the opacity of these “black-box” models, which fail to provide explanations for their actions. All of these scenarios highlight the semantic gap between human and machine analysis, emphasizing the need for human involvement in decision-making as well as the development of Explainable AI tools towards better interpretability of the model.

To mitigate the aforementioned challenges, we propose a novel Explainable Active Learning (XAL) model that combines domain expert assistance and explainable AI (XAI) support within the active learning (AL) paradigm.

In particular, we propose a novel **XAL based semantic segmentation model “SegXAL”** for the driving scene scenarios. *Active learning* facilitates effective training set by iteratively curating the most informative unlabeled data for annotation with the help of human intervention (oracle) accentuating the “human-in-the-loop” paradigm [2],[3]. This “domain expert teaching” emphasizes productivity and enhances trust in AI systems, especially in low-resource as well as high-risk scenarios. Similarly, the *explainability* aspect of the SegXAL model enables the “machine teachers” (human experts) to obtain intuitive reasoning behind the results and to give solicit feedback to the system[4]. This is inspired by the rationale that humans’ cognizance leverages causal and interpretable information to make decisions[5], [6]. Both of these AL and XAI notions within the SegXAL model bridge the semantic gap between man and machine through collaborative intelligence, wherein humans and AI actively enhance each other’s complementary strengths.

The key component of the SegXAL framework is the Explainable Error Mask (EEM) module that provides intuitive reasoning as well as uncertainty measures for the sample selection. The EEM module internally contains two components

viz. Entropy-based Uncertainty (EBU) module and Proximity-aware Explainability (PAE) module. Following popular active learning approaches, the EBU module utilizes uncertainty or disagreement in the unlabeled data to identify the most uncertain and informative samples for annotation by the oracle [7] [8] [9]. Whereas, the PAE module acts as an interpretable proximity approximator that prioritizes the relevant nearby class information, leveraging depth estimation technique and explainable AI. In particular, two advanced AI models viz. MiDaS [10] and DINOV2 [11] are used as the instances of depth estimation. Referring to the XAI technique, we leverage Gradient-weighted Class Activation Mapping (GradCAM) [12], which interprets and visualizes the regions of an input image that are crucial for the network’s prediction of a specific class.

Thus, the PAE module along with the EBU module provides the Explainable-Error Mask (EEM) with both informativeness and explainability, thereby facilitating meaningful annotation from the oracle. Two modes of oracle annotations are presented in this work: The first mode is via **Machine annotated pseudolabels**, wherein the machine itself does an automatic pixel annotation. The second mode is via **Manual annotation**, wherein the human annotator labels the region relevant to the object based on the candidate prompts. The major contributions of the paper are as follows:

- Proposal of a ‘**XAL based semantic segmentation model “SegXAL” for the driving scene scenarios**’, which is the first Explainable Active Learning (XAL) framework in semantic segmentation.
- Development of a novel **Explainable Error Mask (EEM)**, fusing proximity-aware explainability (PAE) and entropy-based uncertainty (EBU) measures, thereby enhancing the efficiency of oracle annotation.
- Proposal of two manual annotations schemes within the Active learning framework viz. **Manual-M** and **Manual-D**, leveraging MiDaS and DINOV2-based explainable error masks, respectively.
- Proposal of a novel **high-confidence sample selection technique based on DICE** similarity coefficient.
- Extensive experimental analysis, ablation studies and state-of-the-art comparative analysis in benchmarking Cityscapes dataset.

The rest of the paper is organized as follows: The related works are described in Section 2. The proposed SegXAL active learning framework is presented in Section 3. The experimental setup and the results are discussed in detail in Section 4 and Section 5 respectively. Finally, the summary of the paper and some future plans are enumerated in Section 6.

2 Related Works

Explainable AI: Explainable Artificial Intelligence (XAI) is an emerging area of research in machine learning [6]. XAI techniques make AI models more interpretable by humans by divulging the hidden “black-box” and providing insights into how the model arrives at a particular decision. Some of the recent research works have been investigating XAI in such cutting-edge areas, e.g. medical domain to find out the feature importance [13] and to visualize the biologically relevant information [14]. Some XAI models were developed for remote sensing

and satellite applications, [15] to analyze synthetic aperture sonar (SAS) data and for Explainable Machine Learning in Satellite Imagery, respectively. The application of XAI approaches in driving scene scenarios is also reported in the recent literature bestowing ideas towards comprehensible and trustworthy autonomous driving technologies [16].

Active Learning: Active Learning (AL) entails the training process of a learning algorithm through an iterative collaboration with a human oracle [17]. AL involves selecting the most relevant data samples from a pool of unlabeled data based on uncertainty, representativeness, or diversity scores computed directly with the model [7],[8]. To this end, some popular approaches to obtain confidence, margin and uncertainty measures are via entropy [18], Softmax probabilities [19], Monte Carlo dropout [20] and Ensemble methods [21]. Such AL models have been widely applied in various vision applications, such as medical scenarios [22], satellite imagery analysis [23] etc. The necessity for AL frameworks for autonomous driving scenarios is reported in [24], mentioning that 'vehicles need 11 billion miles of driving (500 years of nonstop driving with a fleet of 100 cars) to perform just 20 per cent better than a human.' Motivated by this notion, some recent AL works on driving scenes were reported in the literature [25].

AL for semantic segmentation: There are AL methods specially designed for semantic segmentation that work at image, region or pixel levels [7], [8]. The Variational Adversarial Active Learning (VAAL) approach employs adversarial learning to determine whether the latent space signifies labelled or unlabeled data [26]. The work Difficulty-awarE Active Learning (DEAL) [7] incorporates the semantic difficulty to measure the informativeness and select samples at the image level. Another work 'ViewAL' [8] leverages inconsistencies in model predictions across view-points to measure the uncertainty of super-pixels. Yet another work S4AL [27] utilizes pseudo labels generated with a teacher-student framework to identify image regions that help disambiguate confused classes.

Contrary to the aforementioned AL approaches that measure uncertainty/informativeness, our proposed SegXAL additionally augments the notion of explainability in the model. In particular, the PAE module in our proposed SegXAL model imparts contextual and proximity-aware explainability to the oracle to prioritize the annotation of nearby objects, which are pivotal in autonomous driving scenarios. This kind of explainable active learning (XAL) in semantic segmentation is proposed for the first time, to the best of our knowledge. Further, the significance of pixel-level and object-level annotation by the oracle (Machine annotator vs. Human annotator) is also investigated in our proposal.

3 Methodology: SegXAL - Explainable Active Learning for semantic segmentation

The Active Learning (AL) protocol ensures that by intelligently selecting instances for labelling, a learning algorithm can achieve good performance with significantly less training data. Formally, it can be expressed as follows: Let

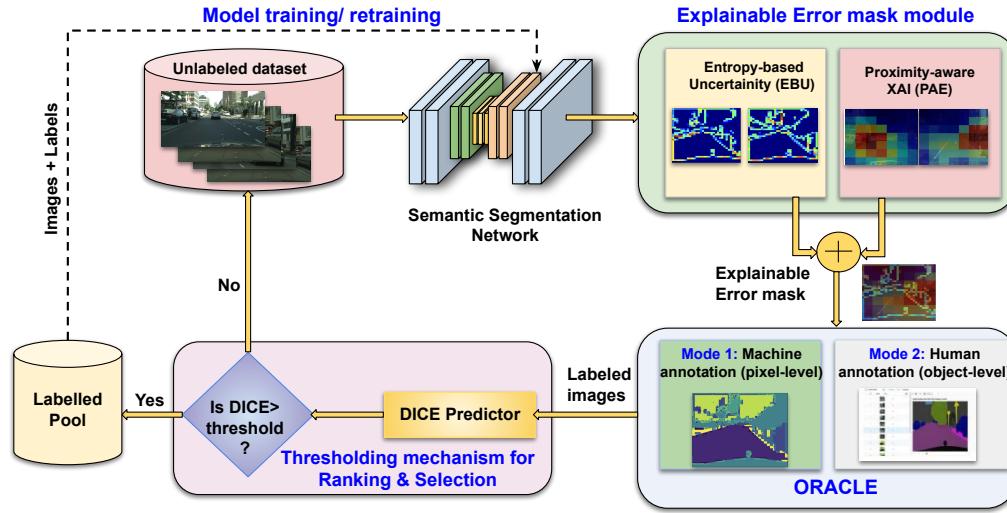


Fig. 1. Visual representation of Explainable Active Learning for semantic segmentation (SegXAL) framework. The framework starts with an initial segmentation of unlabeled data, leveraging pre-trained semantic segmentation deep neural network (e.g. U-net). Further, the Explainable Error Mask (EEM) module computes the uncertainty measure and proximity-aware XAI mask. Based on this EEM output, machine/human expert (oracle) makes intuitive labelling feedback to the system. Further, based on the Dice predictor-based query ranking mechanism, reannotated data are used for labeled pool update and model retraining.

(x^l, y^l) be an annotated sample from the original labelled dataset D^L and x^u represent an unannotated sample from a significantly larger pool of unlabeled data, D^U . The goal of AL is to iteratively query a subset D^S , that contains the most informative n samples $x_1^u, x_2^u, \dots, x_n^u$ from D^U in an iterative manner, given n is the fixed labelling budget.

In this work, we present a novel Explainable Active Learning paradigm for semantic segmentation (SegXAL) in driving scene imagery. Refer to Fig. 1 for the overall architecture of the SegXAL framework. It contains training of the model, prediction of semantic maps, “Explainable Error Mask” (EEM) computation, annotation, selection mechanism and retraining steps. Each of these steps is explained in detail in the forthcoming subsections:

3.1 Step 1: Semantic Segmentation - Training & Prediction

We leverage U-Net [32] as the semantic segmentation network for the model training. Typically, any segmentation model such as FCN [31] or DeepLab [33], among others, could also be utilized. U-Net is employed in this pilot study, due to its ability for the precise localization of objects while maintaining a high level of contextual information as well as lower memory consumption. The U-Net model embodies an encoder-decoder framework. The encoder is responsible for the initial feature extraction and dimensionality reduction, by utilizing successive convolutional and pooling layers followed by nonlinear activation functions

(ReLUs) and batch normalization. Whereas, the decoder works on reconstructing the feature map to the original image size for detailed segmentation using transposed convolutions (or deconvolutions). It also incorporates skip connections, that concatenate feature maps from the contracting path to preserve the high-resolution details that are crucial for accurate segmentation.

In this initial step, a small randomly selected subset of the labelled dataset D_L will be used to train a semantic segmentation network. Following the widely adopted protocol, we randomly sample 10% of the data as labelled data from the train set as our labelled data pool¹. After training the network on D^L , the model performance is evaluated on unlabeled dataset D^U . AL approach strives to forecast which samples from this unlabeled segment of dataset, are most likely to provide the most informative insights, given the current state of the network. To this end, a novel **Explainable Error Mask (EEM)** module is proposed.

3.2 Step 2: Explainable Error mask Module

The Explainable Error Mask (EEM) module is the key component of our SegXAL framework. In contrast to the vanilla Active learning models that provide uncertainty/ representativeness insights for the annotation, this novel EEM module presents an explainable error mask for the interactive annotation by the oracle. It consists of the following components: *i) Entropy-based Uncertainty (EBU), ii) Proximity-aware XAI (PAE) and iii) fusion of PAE and EBU.*

i) Entropy-based Uncertainty (EBU) module:

One of the most important postulations in active learning strategy is to guide the user towards the most relevant areas to annotate, to fix errors. To this end, some standard uncertainty measuring techniques such as entropy [34], or ODIN [35] are exploited in the literature. Following many of the popular AL pipelines, our EBU module leverages entropy metric to measure the uncertainty/disagreement for the unlabeled data, to obtain the most uncertain data which is informative and worthwhile ones to be annotated by the oracle.

Entropy is a measure of uncertainty or information content in a probability distribution [34]. In the context of image segmentation, it is commonly used to quantify the uncertainty of pixel-wise predictions across different classes within a batch of segmented images. Let us denote a batch of segmented images as X with dimensions $[B, C, H, W]$, where B is the batch size, C is the number of classes, H is the height and W is the width of images. Each image in the batch consists of pixel-wise predictions across C classes. The entropy $H(x_{i,j})$ for each pixel $x_{i,j}$ can be calculated as:

$$H(x_{i,j}) = - \sum_{c=1}^C P(c|x_{i,j}) \log_2(P(c|x_{i,j})) \quad (1)$$

where $P(c|x_i)$ represents the probability that pixel $x_{i,j}$ belongs to class c . Higher entropy values indicate greater uncertainty or ambiguity in the predictions, implying lower confidence in the model’s predictions. Conversely, lower entropy values signify higher confidence or clarity in the predictions.

¹ (Ablation studies are carried out by varying the splits of labelled data pool i.e. 10%, 15%, 20%, 25%, 30%, 35%, 40%)

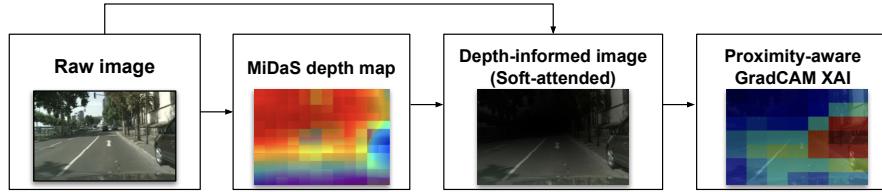


Fig. 2. Proximity-aware Explainable-AI (PAE) Module using MiDaS depth estimation technique. Analogous to MiDaS, DINoV2 depth map is also investigated in this paper.

ii) Proximity-aware Explainable-AI (PAE) module:

The high entropy pixels generated by the Entropy-based Uncertainty (EBU) module can be spread across the entire image, making it challenging from an Oracle perspective to determine where to prioritize attention. Consequently, this may lead to missing out of some of the vital regions to be annotated first. For instance, in driving scene imagery with high entropy scores in the sky, vegetation, and vehicles, annotation priority should be given to nearby classes i.e. vehicles, considering safety concerns. We hypothesize that such a proximity awareness can improve the oracle annotation. In addition, uncertainty techniques often lack human interpretability, hindering an intuitive understanding of why certain regions are crucial for annotation.

Based on the aforesaid rationale, we propose a novel **Proximity-aware Explainable-AI (PAE)** module to mitigate the priority and interpretability concerns. The PAE module is capable of focusing on the key objects and regions of interest in the proximity regions with the help of an explainability heatmap. The working pipeline of our proposed PAE module is depicted in Fig. 2. Either MiDaS or DINoV2 model is leveraged to obtain the given image's relative depth map. MiDaS [10] is a robust monocular depth estimation technique that employs mixed-dataset training to create a robust and generalizable depth estimation model. Whereas, DINoV2 [11] is a self-supervised vision transformer model that uses a teacher-student architecture to provide object-level feature extraction. Both of the models are capable of providing monocular depth map outputs. By integrating the MiDaS/DINO-v2 patchwise depth map with the raw image using a thresholding mechanism, the proximity coverage will be estimated. This results in a depth-informed or soft attention image as shown in Fig. 2. Note that the threshold for generating a depth-informed image varies with each image based on the proximity of the nearest objects. Upon this image, a Gradient-weighted Class Activation Mapping (GradCAM) [12] explainability map is applied to visualize the important objects and regions. GradCAM is a technique for visualizing CNN decisions, highlighting regions crucial for predictions. The mathematical equation for GradCAM activation at spatial position (i, j) for class c i.e. $\text{Grad-CAM}_{i,j}^c$ can be summarized as:

$$\text{GradCAM}_{i,j}^c = \text{ReLU} \left(\sum_k \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial f_k(i,j)} \cdot f_k(i,j) \right), \quad (2)$$

where, y^c is the output score for class c before softmax, $f_k(i,j)$ is the activation

value of the k^{th} feature map at spatial position (i, j) and Z is the normalization constant, typically sum of positive gradients. By applying the GradCAM upon the depth-informed image, we obtain the proximity-aware GradCAM explainability map i.e. $\text{ProxGradCAM}_{i,j}^c$, which prioritizes the object class information which is relevant in the proximity region.

iii) Fusion of PAE and EBU modules:

The PAE heatmap $\text{ProxGradCAM}_{i,j}^c$ is further fused with EBU uncertainty heatmap $H(x_{i,j})$, to obtain the Explainable Error mask $EEM_{i,j}$. Formally,

$$EEM_{i,j} = \alpha \cdot \text{ProxGradCAM}_{i,j}^c + \beta \cdot H(x_{i,j}) \quad (3)$$

where α and β are the weights for the $\text{ProxGrad-CAM}_{i,j}^c$ and $H(x_{i,j})$, respectively. Albeit we used equal contribution for the weights in this work, it can be made learnable.

3.3 Step 3: Oracle for annotation

Next, we acquire labels for the superpixels/Region of Interest (ROI) selected by EEM module, with the help of oracle. In particular, two modes of oracle annotations are envisaged in this work: machine and human oracle. In the former mode (Machine oracle), automatic pixel annotations are simulated by the machine itself. We term these annotations as ‘**pseudolabels**’. In the latter mode (Human oracle), the reannotations are carried out manually by a domain expert. By keeping the interpretable information of the potential error map obtained from EEM as a reference, the annotation process is carried out using tools like Label Studio². Specifically, two manual annotation schemes are devised within the Active learning framework viz. **Manual-M** and **Manual-D**, leveraging MiDaS and DINOv2-based explainable error masks, respectively.

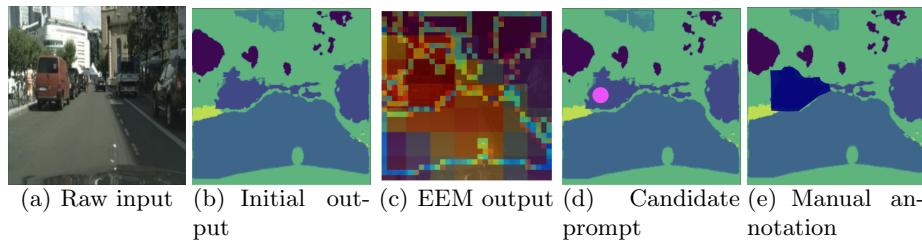


Fig. 3. Oracle’s Reannotation workflow. The magenta point shown in 3(d) is the EEM output prompt corresponding to the relevant object candidate to be annotated.

Fig 3 depicts a sample human oracle-based reannotation workflow. Based on the initial segmentation mask output from the raw image as shown in Fig. 3(b), EEM produces the output $EEM_{i,j}$ (Refer Fig. 3(c)). Further, based on the object candidate prompt as shown 3(d), the human annotator corrects the miss-segmented image regions by providing object-level annotation (Fig. 3(e)). These newly reannotated segmentation masks will be further fed into the sample selection module towards the next iteration of the AL loop.

² Label studio: <https://labelstud.io/>

3.4 Step 4: Thresholding Mechanism for Sample Selection

After the oracle, the labeled images are fed into the Ranking & Selection module. Analogous to the high-confidence sample selection techniques as in [19], we use a novel thresholding mechanism to select high-confidence samples to be incorporated into the labeled data pool. In particular, a standard evaluation metric i.e. ‘DICE predictor’ is utilized to compute the quantitative measure of performance of the segmented images. Mathematically, DICE computation can be written as:

$$\text{DICE} = \frac{2 \times |A \cap B|}{|A| + |B|} \quad (4)$$

where A represents the segmented image and B denotes the reannotated pseudo labels/ human annotations, within each AL cycle.

This **DICE predictor-based sample selection strategy** is devised based on the assumption that “*in every AL cycle, the oracle contributes a significant amount of annotation to improve the quality of semantic segmentation*”. Based on this intuition, we postulate that whenever the similarity between the segmented image and the reannotated image becomes high, a convergence is achieved in the segmentation result. In other words, even after a significant amount of contribution from the oracle, the segmentation result does not improve further, which can be observed as an increase in the DICE similarity coefficient. To guarantee the reliability of high-confidence sample selection, at the end of each iteration, this DICE value is compared against a predefined threshold θ . If the DICE score is above θ , select it and add to the labeled pool and clear it from the unlabeled set; otherwise, feed it back to the unlabeled dataset placed in the unlabelled pool for potential future iterations.

3.5 Step 5: Iterative Active Loop for Semantic Segmentation Improvement

After the Ranking & Selection module, high-confidence segmentation images are added to the labelled data pool D^L , as shown in Fig. 1. Based on this updated dataset, the semantic segmentation model retraining will be carried out. This concludes a complete active learning cycle. Further, a new AL cycle will start based on the updated model weights and the unlabeled dataset D^U . All the series of steps - *Semantic map prediction from unlabelled data, EEM computation, Annotation, Ranking & Selection and Retraining* - are repeated until the labelling budget is reached or all the data is labelled. This iterative AL cycle optimally selects the most informative samples via EEM information and Oracle annotation, enhancing model performance with minimal labelling costs.

4 Experimental Setup

Dataset: We evaluate our proposed SegXAL framework on the Cityscapes dataset for semantic segmentation [30]. Cityscape is a large-scale benchmark for urban street scene understanding, at 1024×2048 pixel resolution with 30 classes including road, car, pedestrian, bicycle, traffic sign, and more. The dataset is divided into three subsets: *train* (2975 images), *validation* (300 images), and

test (500 images). We follow the widely adopted protocol for the dataset - we sample 40% of the data from the trainset as our labelled data pool D^L for initial training then iteratively query 5% new data from the remaining training set, which is used as the unlabeled data pool D^U . Considering samples in the street scenes have high similarities, we first randomly choose a subset D^S from the entire pool of D^U , then query n samples from the subset.

Evaluation protocol: We evaluate our proposed SegXAL model using the standard segmentation evaluation metrics i.e. Intersection over Union (**IoU**) and DICE coefficient. To assess the accuracy of pixel-wise classification, the standard evaluation metric IoU (Intersection over Union) score is utilized. IoU is computed as the ratio of the intersection and union of the ground truth mask and the predicted mask for each class. Further, the DICE similarity coefficient is utilized for ranking & selection of samples, as described in Section 3.4. It provides a balanced measure of segmentation accuracy, especially in cases of class imbalance, and hence is used for our sample selection strategy.

Implementation details: The images with a dimension of 256×512 are normalized using the RGB mean and standard deviation of ImageNet before passing to the network. Our baseline UNet model was evaluated using a stratified K-fold cross-validation approach to ensure robustness and generalizability. The network is trained using a Stochastic Gradient Descent (SGD) optimizer with the following hyper-parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, batch size = 16, initial learning rate = 0.0001. The batch size used is 16 images. For all methods and the upper bound method with the full training data, we train 100 epochs with an unweighted cross-entropy loss function. The proposed method is implemented using the PyTorch framework. The implementation was done in a machine with NVIDIA DGX A100 GPU with 24GB RAM and takes around 8 hours to train the model.

5 Experimental Results

5.1 Evaluation Results

To verify the effectiveness of our proposed SegXAL framework, various quantitative and qualitative analyses are carried out in the Cityscape dataset. The mean Intersection over Union (mIoU) at each AL stage i.e. 10%, 15%, 20%, 25%, 30%, 35%, 40% of the full training set are adopted as the evaluation metric. Every method is run 5 times and the average mIoUs are reported.

Refer to Table 1 for the per-class IoU and mIoU for each method at the fifth AL cycle, using 40% training data in the Cityscapes dataset. Compared to other popular approaches such as DEAL[7] and Core-set[36], SegXAL is found to be outperforming in overall mIoU (Pseudolabels-63.56; Manual-M -64.37; Manual-D -65.11), as well as on various classes, such as road, building, wall, traffic light, traffic sign, vegetation, terrain, sky, rider, car and truck. Furthermore, between the two modes of oracle annotation i.e. Pseudolabel vs Manual, we observe that the manual mode outperforms with a 0.8% increase against the former, and has a significant boost in class-wise IoUs. We also provide a statistical measure of standard deviation (STD) to give an insight into the variability of the model

Table 1. Class-wise IoU and mIoU on Cityscape dataset with 40% training data. For clarity, only the average of 5 AL runs are reported, and the best and the second best results are highlighted in **bold** and *italics*.

Method	Road	Sidewalk	Building	wall	Fence	Pole	Traffic	Light	Traffic	sign	Vegetation	Terrain
Fully-supervised	97.58	80.55	88.43	51.22	47.61	35.19	42.19	56.79	89.41	60.22		
Random [27]	96.03	72.36	86.79	43.56	44.22	36.99	35.28	53.87	86.91	54.58		
Entropy [27]	96.28	73.31	87.13	43.82	43.87	<i>38.10</i>	37.74	55.39	87.52	53.68		
Core-Set[36]	96.12	72.76	87.03	44.86	45.86	35.84	34.81	53.07	87.18	53.49		
DEAL [7]	95.89	71.69	87.09	45.61	44.94	38.29	36.51	55.47	87.53	56.90		
Ours (Pseudolabels)	96.67	72.42	87.04	<i>46.91</i>	45.02	36.26	<i>37.83</i>	56.11	<i>87.93</i>	57.54		
Ours (Manual-M)	96.91	72.68	87.44	46.62	45.22	35.62	36.24	55.78	87.66	57.86		
Ours (Manual-D)	96.98	73.43	88.34	46.88	45.38	36.12	37.36	55.38	87.84	59.87		
Method	Sky	Pedestrian	Rider	Car	Truck	Bus	Train	Motor	Cycle	Bicycle	mIoU	STD
Fully-supervised	92.69	65.12	37.32	90.67	66.24	71.84	63.84	42.35	61.84	65.30	19.48	
Random [27]	91.47	62.74	37.51	88.05	56.64	61.00	43.69	30.58	55.67	59.00	20.61	
Entropy [27]	92.05	63.96	34.44	88.38	59.38	64.64	50.80	36.13	57.10	61.46	20.14	
Core-Set[36]	91.89	62.48	36.28	87.63	57.25	<i>67.02</i>	56.59	29.34	53.56	60.69	20.61	
DEAL [7]	91.78	64.25	39.77	88.11	56.87	64.46	50.39	38.92	56.59	61.64	19.41	
Ours (Pseudolabels)	92.18	62.53	38.82	<i>88.61</i>	59.07	65.72	47.12	35.41	55.83	63.56	20.12	
Ours (Manual-M)	92.84	62.73	<i>39.34</i>	87.97	59.43	66.01	46.92	34.98	54.93	64.37	19.96	
Ours (Manual-D)	92.93	62.56	<i>39.07</i>	88.11	<i>59.47</i>	65.70	46.88	35.53	54.71	<i>65.11</i>	20.15	

Table 2. Comparison of mean IoU of SegXAL model over 5 active learning cycles, with 40% training data, using pseudo labels annotated by machine vs and human annotations using MiDaS (Manual-M) and DINOV2 (Manual-D) variants.

Mode	ALcycle1	ALcycle2	ALcycle3	ALcycle4	ALcycle5
Pseudolabel	20.71	27.11	39.23	50.47	63.56
Manual-M	23.62	28.02	39.11	51.33	64.37
Manual-D	24.24	30.02	39.96	52.31	65.11

performance. Further, Table 2 displays the incremental trend of mIoU values over multiple iterations. It is observed that at the end of 5 AL cycles itself, mIoU is improved from 20.71 to 63.56 using Pseudolabels, 23.62 to 64.37 using Manual-M and 24.24 to 65.11 using Manual-D.

5.2 Visualisation results

To demonstrate the efficacy of our proposed EEM module, we visualize the qualitative results. Referring to Fig 4, the visualization of 5 AL cycles of a sample raw image shown in Fig. 2 are depicted column-wise. The pixel entropy, Explainable Error Mask (EEM) output and the machine annotated pseudolabel-based segmentation results are shown along the first, second and third rows respectively.

Referring to Fig. 4 (a)-(e), high entropy areas represented in red or orange patches indicate a high degree of variability in pixel values. Conversely, low entropy regions, in blue, signify homogenous or less complex segments, where pixel intensities are similar and are in their class boundaries. This entropy map thus serves as a useful visualisation tool to analyse the complexity of the scenes over the loops. Further, Proximity-aware GradCAM-XAI is fused with this entropy mask to obtain an Explainable Error Mask as shown in Fig. 4 (f)-(j)(Refer Sec.

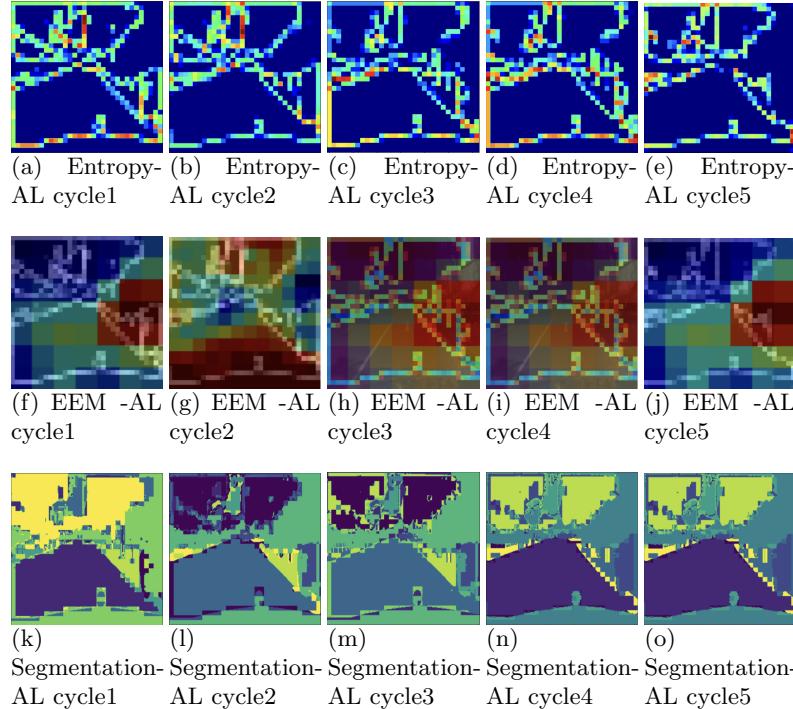


Fig. 4. Visualization of model performance over 5 Active Learning cycles.

3.2). These EEM outputs clearly “*explain*” the oracle to focus and prioritise the annotation of the closer objects/regions with *high entropy*, which are quite critical in decision-making in the real-world scenario. The oracle-annotated results Fig. 4 (k)-(o) depicts the significant improvement in segmentation quality over 5 active learning cycles.

5.3 Ablation Study

i) Impact of Machine based pseudo label annotation vs Manual Annotation/ Impact of Pixel level strategy and object level strategy:

In this ablation study, we analyse the effect of Machine-based pseudo label reannotation and Manual reannotation. As mentioned earlier, the machine oracle mode leverages pixel-level pseudolabel values for annotation whereas the human oracle employs object-level annotation via Label Studio. We could observe from Table 1, Table 2 and Fig. 5 that both approaches provide superior performance in semantic segmentation. Specifically, the manual annotation outperforms the Pseudolabel annotation (Refer Table 1) and smooth segmentation masks (See Fig. 5). Nevertheless, Machine-based auto labelling is faster and bestows a promising automated AL solution from a practical perspective compared to manual annotation, wherein a human expert reviews every image and reannotates.

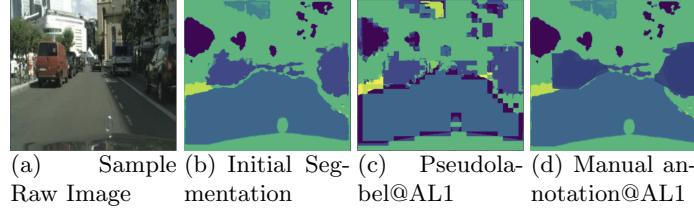


Fig. 5. Visualization of Machine-based Pseudolabel vs. Manual annotation outputs

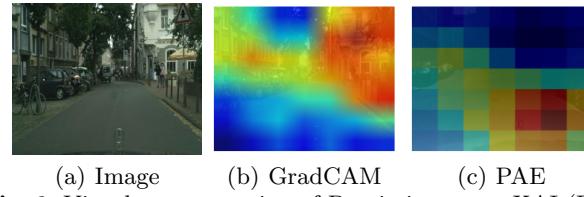


Fig. 6. Visual representation of Proximity-aware XAI (PAE)

ii) Impact of Proximity-aware XAI EBU and PAE modules :

To understand the impact of EBU and PAE modules, quantitative ablation studies are carried out. Referring to Table 3, it can be observed that the lack of EBU sub-module within the EEM block results in a mIoU drop of 3.69, 3.94 and 4.02 in Pseudolabel, Manual-M and Manual-D cases, respectively. Its counterpart results in the absence of PAE sub-modules are 3.47, 2.69 and 3.4 respectively. Additionally, a qualitative study is also conducted to comprehend the visual interpretation of Proximity-aware XAI, as depicted in Fig.6. It is observed that PAE outperforms the Vanilla GradCAM[12], which provides insights of the scene by localizing on the key areas semantic classes via saliency heat maps (Refer Fig.6(a, b)). Built on top of this Grad-CAM concept, our Proximity-aware XAI module refines the attention further onto the nearby objects in the proximity regions e.g. nearby vehicles and sidewalks, as shown in Fig.6(c). This PAE enhancement notably fosters safety and transparency in autonomous driving scenarios.

Table 3. A quantitative study on impact of EEM module and their components.

Mode	PseudoLabels	Manual-M	Manual-D
With EEM	63.56	64.37	65.11
Without EBU	59.87	60.43	61.09
Without PAE	60.09	61.68	61.87

iii) Impact of change in % of data split In this ablation study, we investigate the effect of data split on the SegXAL performance. In particular, we perform various splits of 10%, 15%, 20%, 25%, 30%, 35%, and 40% of the dataset for the initial model training. Referring to Fig. 7 showing the fifth AL cycle mIoU result, it can be observed that based on the increase of labelled data from 10% to 40%, there is a significant increase in mIoU for Pseudolabel 51.02 to 63.56, Manual-M 52.29 to 64.37 and Manual-D 52.83 to 65.11.

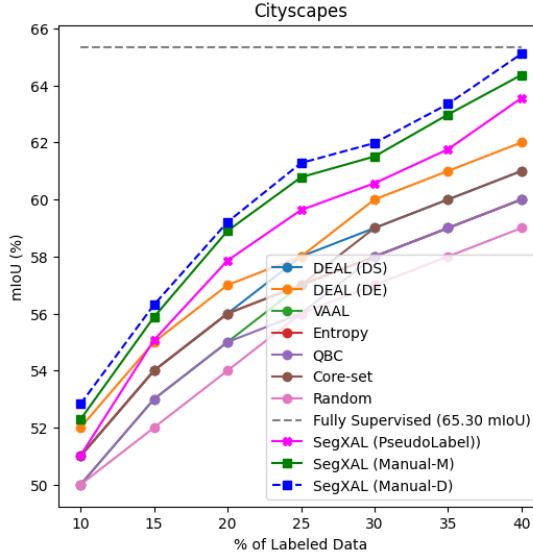


Fig. 7. SegXAL performance against state-of-the-art on the Cityscapes dataset with 40% training data. Every method is evaluated at the end of 5 AL cycles.

5.4 State-of-the-art Comparison

We compare SegXAL with other Active Learning-based semantic segmentation approaches that are deployed on the Cityscapes dataset under similar conditions (with 40% training data over 5 AL cycles) i.e. DEAL[7], core-set approach[36], random, entropy[7,27] and QBC [7]. Although another recent study S4AL[27] achieves a competitive result of mIoU 64.80, it is not included in the comparison due to its different setting of 16% training data. Referring to the results as shown in Table 1 and Fig. 7, it can be observed that SegXAL outperforms the state-of-the-art approaches with a significant margin, achieving the best result of 65.11 mIoU with human annotations with DINOv2 depth map (blue-dotted line). It is also observed from Table 1 that, the segmentation performance on the nearby classes such as road (96.98), sidewalk (73.43), wall (73.48) and vehicles such as truck (59.47), rider (39.34) are better or on par with the previously proposed methods. This superior performance could be accredited to the Explainable Error Mask module that facilitates object-level proximity mechanism using XAI attention and Entropy metric, which prioritizes the highly informative nearby objects’ annotations compared to far away objects such as train, vegetation etc.

6 Conclusions and Future work

In this work, we proposed a novel Explainable Active Learning framework viz. SegXAL for semantic segmentation. A pilot study on the application of the SegXAL model for driving scene semantic segmentation is presented in this paper. In contrast to most of the existing Active learning methods that annotate using uncertainty information, the proposed model additionally “explains” the proximity region of interests and key objects to be prioritized while annotating by the oracle, with the help of a newly proposed Explainable Error Mask (EEM)

module. Such XAI heatmap explanations not only improve the segmentation accuracy but also bridge the semantic gap that exists between human and machine interpretation. Our SegXAL model outperforms state-of-the-art results. Future improvements can be made by introducing better attention mechanisms such as Vision transformers and extending the applications to other driving datasets and other domains.

References

1. T. Talaei Khoei, H. Ould Slimane, and N. Kaabouch, “Deep learning: systematic review, models, challenges, and research directions,” *Neural Computing and Applications*, vol. 35, 09 2023.
2. D. D. Lewis and J. Catlett, “Heterogeneous uncertainty sampling for supervised learning,” in *Machine learning proceedings 1994*. Elsevier, 1994, pp. 148–156.
3. K. Margatina, G. Vernikos, L. Barrault, and N. Aletras, “Active learning by acquiring contrastive examples,” *arXiv preprint arXiv:2109.03764*, 2021.
4. A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai,” *Information fusion*, vol. 58, pp. 82–115, 2020.
5. B. M. Rottman and R. Hastie, “Reasoning about causal relationships: Inferences on causal networks.” *Psychological bulletin*, vol. 140, no. 1, p. 109, 2014.
6. S. C.-H. Yang, N. E. T. Folke, and P. Shafto, “A psychological theory of explainability,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 25 007–25 021.
7. S. Xie, Z. Feng, Y. Chen, S. Sun, C. Ma, and M. Song, “Deal: Difficulty-aware active learning for semantic segmentation,” in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.
8. Y. Siddiqui, J. Valentin, and M. Nießner, “Viewal: Active learning with viewpoint entropy for semantic segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9433–9443.
9. G. Lenczner, A. Chan-Hon-Tong, B. Le Saux, N. Luminari, and G. Le Besnerais, “Dial: Deep interactive and active learning for semantic segmentation in remote sensing,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 3376–3389, 2022.
10. R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.
11. M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, “Dinov2: Learning robust visual features without supervision,” *arXiv preprint arXiv:2304.07193*, 2023.
12. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
13. J. Zuallaert, F. Godin, M. Kim, A. Soete, Y. Saeys, and W. De Neve, “Splicerover: interpretable convolutional neural networks for improved splice site prediction,” *Bioinformatics*, vol. 34, no. 24, pp. 4180–4188, 2018.

14. P. Rajpurkar, A. Park, J. Irvin, C. Chute, M. Bereket, D. Mastrodirosa, C. P. Langlotz, M. P. Lungren, A. Y. Ng, and B. N. Patel, “Appendixnet: deep learning for diagnosis of appendicitis from a small dataset of ct exams using video pretraining,” *Scientific reports*, vol. 10, no. 1, p. 3958, 2020.
15. T. Stomberg, I. Weber, M. Schmitt, and R. Roscher, “Jungle-net: Using explainable machine learning to gain new insights into the appearance of wilderness in satellite imagery,” *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 3, pp. 317–324, 2021.
16. S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, “Towards safe, explainable, and regulated autonomous driving,” in *Explainable Artificial Intelligence for Intelligent Transportation Systems*. CRC Press, 2021, pp. 32–52.
17. B. Settles, “Active learning literature survey,” 2009.
18. B. Settles and M. Craven, “An analysis of active learning strategies for sequence labeling tasks,” in *proceedings of the 2008 conference on empirical methods in natural language processing*, 2008, pp. 1070–1079.
19. K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, “Cost-effective active learning for deep image classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2016.
20. Y. Gal, R. Islam, and Z. Ghahramani, “Deep bayesian active learning with image data,” in *International conference on machine learning*. PMLR, 2017, pp. 1183–1192.
21. W. H. Beluch, T. Genewein, A. Nürnberg, and J. M. Köhler, “The power of ensembles for active learning in image classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9368–9377.
22. A. Liebgott, T. Küstner, S. Gatidis, F. Schick, and B. Yang, “Active learning for magnetic resonance image quality assessment,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 922–926.
23. A. Goupilleau, T. Ceillier, and M.-C. Corbineau, “Active learning for object detection in high-resolution satellite images,” *arXiv preprint arXiv:2101.02480*, 2021.
24. D. Shapiro, “What is active learning?” 2020, nVIDIA blog, [Online]. Available: <https://blogs.nvidia.com/blog/what-is-active-learning/>. [Online]. Available: <https://blogs.nvidia.com/blog/what-is-active-learning/>
25. S. Schmidt, Q. Rao, J. Tatsch, and A. Knoll, “Advanced active learning strategies for object detection,” in *2020 IEEE intelligent vehicles symposium (IV)*. IEEE, 2020, pp. 871–876.
26. S. Sinha, S. Ebrahimi, and T. Darrell, “Variational adversarial active learning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5972–5981.
27. A. Rangnekar, C. Kanan, and M. Hoffman, “Semantic segmentation with active semi-supervised learning,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5966–5977.
28. S. Ruder, “An overview of gradient descent optimization algorithms,” *arXiv preprint arXiv:1609.04747*, 2016.
29. S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, “Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions,” *IEEE Access*, 2024.
30. 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

31. J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
32. O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention-MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.
33. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
34. C. E. Shannon, “A mathematical theory of communication,” *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
35. S. Liang, Y. Li, and R. Srikant, “Principled detection of out-of-distribution examples in neural networks,” *CoRR, abs/1706.02690*, vol. 1, 2017.
36. O. Sener and S. Savarese, “Active learning for convolutional neural networks: A core-set approach,” *arXiv preprint arXiv:1708.00489*, 2017.