

MULTI-TASK PERCEPTION IN UNSTRUCTURED ENVIRONMENTS: ANTI-DEGRADATION COMPLEMENTARY LEARNING AND SAMEHANCER

Anonymous authors

Paper under double-blind review

ABSTRACT

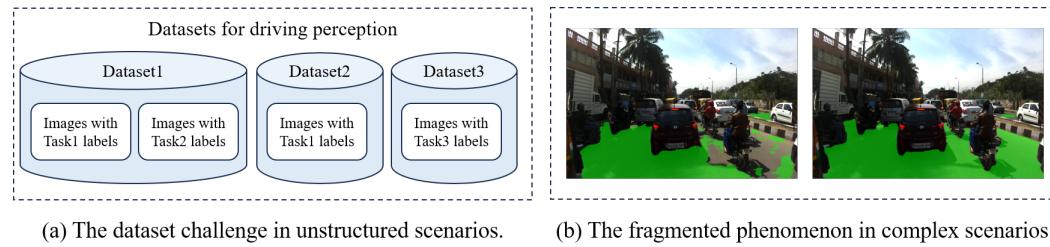
While autonomous driving perception has advanced significantly in structured environments, unstructured environments still present major challenges due to the complexity of traffic participants and irregular road conditions. This paper focuses on addressing these challenges through multi-task perception, targeting drivable area segmentation and object detection in unstructured environments. A key issue in existing datasets for unstructured settings is the non-overlapping annotation of images across different tasks, which limits the efficiency of data utilization. To tackle this, we propose Anti-Degradation Complementary Learning (ADC learning), a semi-supervised approach that allows different tasks to share knowledge across unlabeled data, thereby maximizing the use of available image information. Additionally, we introduce SAMEnhancer, which integrates the Segment Anything Model (SAM) to improve segmentation quality by combining the semantic specificity of network training with the coherence of SAM’s segmentation. Extensive experiments validate the effectiveness of our methods, demonstrating significant performance improvements in both segmentation and detection, especially in challenging unstructured scenarios.

1 INTRODUCTION

Autonomous driving technology has made significant strides in recent years, particularly in structured environments such as highways and urban roads. In these environments, the presence of clear lane markings, predictable traffic patterns, and standardized signage has allowed perception systems to achieve high levels of accuracy in tasks like lane detection, drivable area segmentation, and object detection. However, the transition to unstructured environments—characterized by irregular road layouts, diverse terrains, and a wide range of unpredictable traffic participants—presents a new set of challenges that current perception systems are not fully equipped to handle.

One of the critical challenges in advancing perception for unstructured environments is the availability of suitable datasets. While datasets like the Indian Driving Dataset (IDD) offer a rich array of images from unstructured scenarios, they often suffer from a lack of consistency in annotations across different tasks. Specifically, in the IDD dataset, the images annotated for semantic segmentation are not the same as those annotated for object detection. This non-overlapping nature of annotations reduces the efficiency of data utilization, as it limits the ability to jointly optimize multiple perception tasks using a single dataset. Additionally, there are several smaller datasets that are not as extensive as IDD, which collect data from unstructured areas such as rural and wilderness regions. While these datasets exhibit similar distributions, some contain segmentation annotations, while others include labels for depth estimation and object detection, illustrated in Figure 1, (a). The challenge lies in how to integrate these smaller datasets to enable mutual learning, allowing them to benefit from each other’s labels and enhance their respective perceptual capabilities. Inspired by this challenge, this paper introduces the concept of Anti-Degradation Complementary Learning (ADC learning), a multi-stage semi-supervised multi-task training strategy that allows for the separate yet coordinated training of multi-tasks, thereby maximizing the use of available data.

Meanwhile, in unstructured environments, the complexity of the scene increases drastically. Roads may lack clear boundaries, obstacles can appear unexpectedly, and the types of objects encountered



054
055
056
057
058
059
060
061
062
063
064
065
066
067
Figure 1: The challenges in unstructured scenarios. (a) demonstrates the dataset challenge, and (b)
shows the fragmented phenomenon, where the left side displays the predicted result and the right
side displays the result processed by SAMEnhancer.

068
069
070
071
072
073
074
075
076
077
078
079
can vary widely. These factors complicate the perception tasks that are relatively straightforward in
structured settings, necessitating more advanced and adaptable models. Moreover, the diversity and
irregularity of unstructured environments mean that perception systems must be capable of handling
a much broader range of scenarios, further increasing the difficulty of developing robust and reliable
models. Segmentation based on densely annotated pixel points often lacks sufficient adaptability,
resulting in fragmented representations of the road surface in complex environments, as shown in
Figure 1, (b). To enhance the robustness and integrity of road recognition in unstructured scenes,
this paper introduces the SAMEnhancer tool. SAM, derived from self-supervised training, possesses
strong adaptability and generalization capabilities, along with a robust structural understanding of
the data itself. SAMEnhancer utilizes a lightweight model, Mobile SAM, to integrate the strong
semantic recognition of the original results with the robust structural properties of SAM, creating a
plug-and-play lightweight tool that ensures structural integrity in recognition results within complex
scenes.

080 Our contributions can be summarized as follows:
081

- We propose the Anti-Degradation Complementary learning (ADC learning) strategy that addresses the challenge of non-overlapping task annotations, enabling more efficient data utilization, especially in scenarios where the data is insufficient.
- We design a plug-and-play tool, SAMEnhancer, that leverages SAM to enhance the coherence and integrity of segmentation in complex scenes.
- We demonstrate the effectiveness of our approach through extensive experiments, showing significant improvements in drivable area segmentation and object detection in challenging unstructured scenarios.

091 2 RELATED WORK 092

093 2.1 MULTI-TASK LEARNING FOR PERCEPTION 094

095 Multi-task learning (MTL) has become a pivotal approach in computer vision, enabling the simultaneous
096 learning of multiple tasks within a single framework. This strategy is particularly effective in
097 perception systems, where tasks such as object detection, semantic segmentation, and depth estimation
098 often share underlying features. Early work in MTL leveraged hard parameter sharing, where
099 different tasks shared the same network backbone while maintaining task-specific output layers.
100 Subsequent research explored soft parameter sharing, allowing each task to have its own set of pa-
101 rameters while still enabling information exchange across tasks, thereby mitigating task interference
102 and enhancing learning flexibility (Lee & Liu, 2021).

103 Recent advancements have seen MTL applied to autonomous driving, where perception tasks must
104 be performed in real-time and in diverse environments. LSNetDuan et al. (2021) consolidates ob-
105 ject detection, instance segmentation, and pose estimation under the umbrella of location-sensitive
106 visual recognition, employing a unified approach to tackle these tasks. Furthermore, Phillips et al.
107 Phillips et al. (2021) explored deep multi-task learning for joint localization, perception, and prediction,
highlighting the efficiency and accuracy gains achieved in complex real-world driving scenar-

108 ios. DLT-NetQian et al. (2019) adopts the encoder-decoder framework and actively constructs con-
 109 text tensors between subtask decoders, facilitating the sharing of specific information across tasks.
 110 MultiNetTeichmann et al. (2018) utilizes a shared encoder alongside three independent decoders to
 111 simultaneously address the three scene perception tasks: scene classification, object detection, and
 112 segmentation of the drivable area. In Wu et al. (2022), the YOLOP framework demonstrated the ef-
 113 fectiveness of MTL in reducing computational costs and improving accuracy through an end-to-end
 114 network capable of simultaneously performing lane detection, object detection, and drivable area
 115 segmentation. This integration streamlined the processing pipeline and enhanced overall system
 116 performance. Their approach aimed to create more robust models by sharing knowledge between
 117 tasks such as object detection, semantic segmentation, and depth estimation. Lastly, the FULLER
 118 framework addressed limitations in existing multi-modal MTL methods in 3D autonomous driving
 119 scenarios, unifying multi-modality and multi-task learning through a multi-level approach, which is
 120 crucial for managing diverse data sources and tasks in autonomous driving Huang et al. (2023).

121 However, these approaches typically rely on well-aligned datasets with consistent annotations across
 122 all tasks, limiting their applicability in more complex scenarios like unstructured environments.

124 2.2 DATASETS FOR UNSTRUCTURED ENVIRONMENTS

126 To develop perception models suited for unstructured environments, it is essential to use datasets
 127 that accurately reflect the unique challenges of these settings. One such dataset is the Robot Un-
 128 structured Ground Driving (RUGD), which comprises over 7,000 RGB frames with pixel-wise an-
 129 notations for semantic segmentation in off-road environments. Captured from a small unmanned
 130 mobile robot, RUGD presents challenging visual properties, including blurred frames and irregular
 131 class boundaries Wigness et al. (2019). Another noteworthy dataset is RELLIS-3D, collected on the
 132 Rellis Campus of Texas A&M University, which includes 13,556 LiDAR scans and 6,235 images
 133 across five traversal sequences on non-paved trails, documenting diverse environments like forests
 134 and pastures Duraisamy & Natarajan (2023). The ORFD dataset focuses on free space detection in
 135 various off-road scenes, such as forests and farmlands, under changing weather and lighting condi-
 136 tions Chang et al. (2007). The Freiburg Forest dataset, collected using a Viona autonomous mobile
 137 robot, features multi-spectral and multi-modal images at 20 Hz with a resolution of 1024x768 pixels,
 138 providing annotated pixel-wise segmentation masks for six classes Valada et al. (2017).

139 Among these datasets, the Indian Driving Dataset (IDD) Varma et al. (2019) stands out as one of the
 140 most comprehensive, capturing a wide array of scenarios, including rural roads, crowded city streets,
 141 and diverse weather conditions. This dataset provides annotations for semantic segmentation, ob-
 142 ject detection, and instance segmentation, making it a valuable resource for multi-task learning in
 143 unstructured environments. Data collection for IDD was conducted in India, where road conditions
 144 differ significantly from those in Europe and North America, featuring a diverse array of traffic
 145 participants, such as autorickshaws and animals. Compared to Cityscapes, IDD exhibits a notably
 146 different class distribution, with a higher proportion of motorcycles and two-wheeled vehicles. The
 147 diversity of background classes and ambient factors further contributes to the complexity of road
 148 scenes. While IDD offers 10,004 labeled images with finely delineated instance-level boundaries,
 149 its annotation types surpass those of Cityscapes in terms of object classes and appearance diversity.

150 Research by Baheti et al. Baheti et al. (2020) focuses on IDD, proposing modifications to the
 151 DeepLab V3+ Chen et al. (2018) framework by using lower atrous rates to improve dense traffic
 152 prediction, with a dilated Xception network serving as the backbone for feature extraction. These
 153 explorations lay the groundwork for achieving higher performance on the IDD dataset.

154 However, a notable limitation of the IDD dataset is the lack of consistency in annotations across
 155 tasks. Specifically, the segmentation and detection annotations are not provided for the same set of
 156 images, which complicates the training of joint models. This issue has prompted the need for novel
 157 training strategies, such as the alternating multi-task training proposed in this work, to fully exploit
 158 the potential of the IDD dataset.

159 2.3 SEGMENT ANYTHING MODEL (SAM) AND ITS APPLICATIONS

160 The Segment Anything Model (SAM) represents a significant leap forward in the field of image seg-
 161 mentation. Designed as a universal segmentation model, SAM can generate high-quality masks for

162 a wide variety of objects in diverse scenarios without the need for fine-tuning on specific datasets.
 163 It operates using a set of point prompts that guide the segmentation process, making it highly adapt-
 164 able and effective in situations where traditional segmentation models may struggle. Recent ad-
 165 vancements in SAM are highlighted in several studies, including the work by Kirillov et al. (2023),
 166 which emphasizes its capability to generate high-quality segmentation masks across various contexts
 167 Kirillov et al. (2023).

168 SAM has been successfully applied in various domains, from medical imaging to autonomous driv-
 169 ing, where it has enhanced the precision of segmentation tasks. In the context of multi-task learning,
 170 SAM’s ability to generate accurate masks from minimal input has the potential to significantly im-
 171 prove the performance of downstream tasks, such as object detection and segmentation. Zhang et al.
 172 (2023) provide a comprehensive survey on SAM, exploring its implications for vision and beyond,
 173 underscoring its versatility Zhang et al. (2023b). In this work, we integrate SAM into our multi-task
 174 perception framework, using it to generate masks that are fed into both segmentation and detection
 175 branches, thereby improving overall accuracy. Setu et al. (2024) further unveil SAM’s functionality,
 176 applications, and practical implementation across multiple domains Setu et al. (2024).

177 However, SAM’s high computational complexity leads to suboptimal performance in real-time ap-
 178 plications. To address this issue, researchers have proposed a series of acceleration and optimization
 179 methods. First, some studies have introduced lightweight encoders and decoupled distillation tech-
 180 niques to significantly reduce the model’s computational complexity and size, enabling fast and
 181 efficient image segmentation on mobile devices Zhang et al. (2023a). Additionally, other research
 182 has proposed new training strategies and hardware acceleration schemes, greatly enhancing segmen-
 183 tation speed to meet the demands of real-time applications Zhao et al. (2023). Furthermore, some
 184 studies have designed efficient visual transformer architectures that optimize model structure and
 185 parameter configurations, significantly increasing processing speed without compromising perfor-
 186 mance, making them suitable for resource-constrained environments Zhang et al. (2024).

187 These studies demonstrate that through optimization and acceleration techniques, SAM can achieve
 188 efficient operation in real-time applications while maintaining high accuracy. In summary, the pow-
 189 erful object generalization ability of SAM, combined with these optimization methods, provides
 190 robust support for feature representation and perception in complex scenes, thereby enhancing the
 191 overall effectiveness of perception models across diverse applications.

193 3 METHOD

196 This paper uses YOLOP as the foundational structure. Due to the diverse types of unstructured
 197 roads, which include dirt roads, concrete roads, and sandy roads, and the fact that lane markings
 198 may not always be present on the surface, so we remove the lane detection branch from YOLOP
 199 while retaining the branches for object detection and road segmentation. To enhance the utilization
 200 of unstructured road data, we propose a complementary learning approach that links two originally
 201 independent datasets for multi-task learning. To improve the accuracy and coherence of road seg-
 202 mentation, we introduce SAMEnhancer, which leverages the advantages of self-supervised learning
 203 to enhance segmentation results. The following sections will provide a detailed introduction.

205 3.1 ANTI-DEGRADATION COMPLEMENTARY LEARNING

207 In structured scenarios, a commonly employed strategy is multi-task learning techniques, as illus-
 208 trated in Figure 2 (a). This approach aims to achieve the simultaneous execution of various tasks
 209 within a unified framework using different task data annotations based on the same dataset, thereby
 210 enhancing perceptual robustness and generalization. However, the data in unstructured scenarios
 211 is not as comprehensive as that in structured scenarios, and many datasets can only support single
 212 tasks. For example, R2D2 includes labels for object detection, ORFD contains labels for segmenta-
 213 tion, and IDD features labels for image segmentation and localization, with no overlapping images
 214 corresponding to the labels. This leads to the following issues:

- 215 1. Low Data Utilization: A single image may contain information for all perceptual tasks, but
 insufficient annotations limit the extractable information from that image.

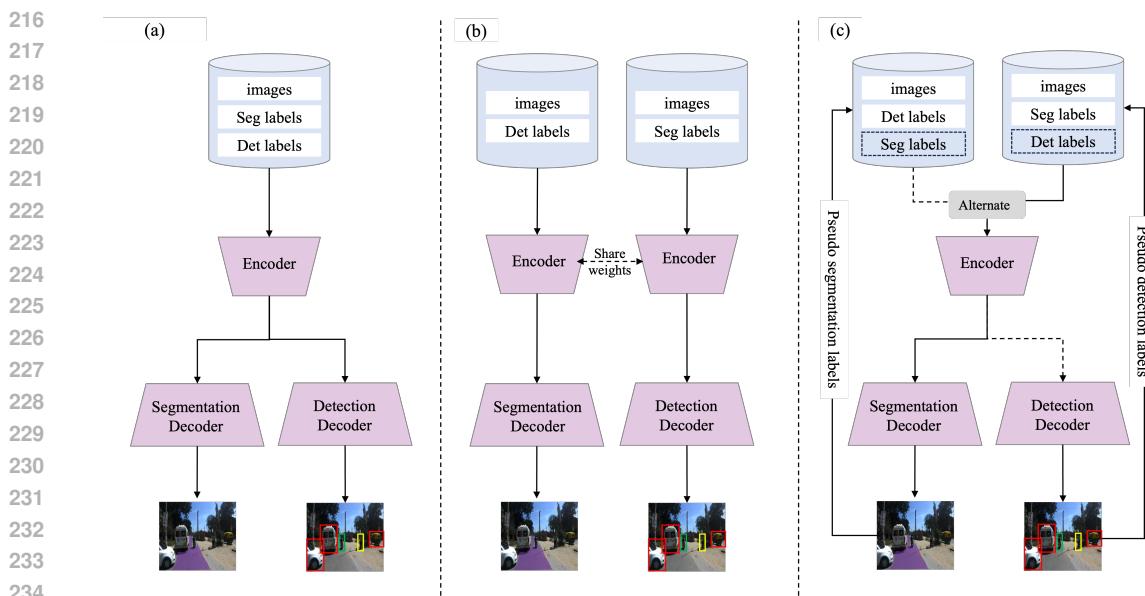


Figure 2: The contrast of different training strategies. (a) represents a multi-task training network with complete data labels, (b) illustrates collaborative training of two tasks through a shared encoder when data labels are incomplete, achieving the goal of multi-task learning, and (c) depicts the training strategy proposed in this study, where alternating training allows the two tasks to provide each other with pseudo-labels, realizing Anti-Degradation Complementary Learning.

2. Limited Data for Individual Tasks: The current scale of perception data in unstructured scenarios is relatively small, providing limited data for single tasks.

A straightforward strategy is to have the network alternately learn on two tasks, directly sharing weights in the feature extractor to enhance the network's cross-task capabilities, as depicted in Figure 2 (b). However, this strategy often encounters network degradation issues, meaning that after training on the next task, there is a significant drop in performance on the previous task.

To address the existing problems, this paper proposes a method called Anti-Degradation Complementary Learning (ADC learning) based on a multi-task learning framework, incorporating alternating learning strategies and semi-supervised learning, as shown in Figure 2 (c). ADC learning consists of two phases: Phase One involves simple alternating training as illustrated in Figure 2 (b), utilizing a shared encoder trained on their respective datasets. When training one task, the branch for the other task is frozen. This learning process runs for a limited number of epochs to acquire more generalized knowledge from the two tasks. Phase Two activates the pseudo-label training module, where in each training stage, one task is trained using ground truth labels while the other utilizes pseudo-labels. For each training stage involving pseudo-label training, the pseudo-labels are generated from the model trained on the ground truth in the previous phase, applied to the current task's dataset.

With the intervention of pseudo-label supervision, the network can continue learning the knowledge of the current task while retaining the knowledge gained in the previous phase, effectively preventing model degradation. This creates a semi-supervised multi-task complementary learning state where originally independent datasets can leverage each other's information to accomplish perceptual tasks that lack ground truth. This significantly enhances the utilization of data and labels while improving model generalization.

3.2 SAMENHANCER

The network training for the target road enables the model to have pixel-wise semantic discrimination capabilities, which, while sufficiently fine-grained, also presents an issue of inconsistency: pixels with the same semantics are often concentrated in certain regions, and pixel-level discrim-

270 /* alternate training stage: */
 271 **begin**
 272 mode = segmentation or detection;
 273 **while** epoch < mix_epoch **do**
 274 **while** i < alternate_frequency **do**
 275 freeze !mode decoder;
 276 input mode data;
 277 model.train();
 278 compute supervised loss;
 279 **end while**
 280 switch mode;
 281 **end while**
 282 **end**
 283 /* Mix training stage: */
 284 **begin**
 285 unfreeze all layers;
 286 **while** epoch < end_epoch **do**
 287 **if** !epoch % update_frequency **then**
 288 | update pseudo labels
 289 **end if**
 290 input dataset with pseudo labels;
 291 model.train();
 292 compute supervised and semi-supervised loss;
 293 **end while**
 294 **end**

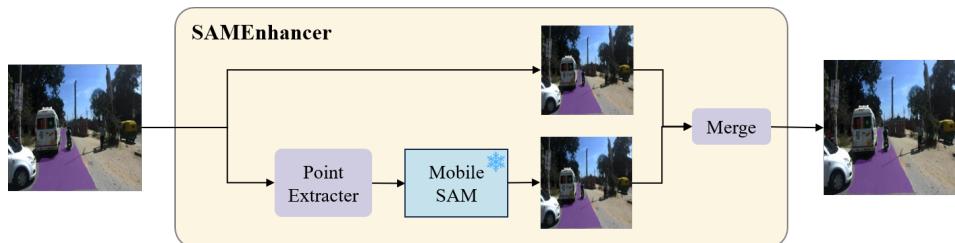


Figure 3: The process of the proposed SAMEnhancer algorithm.

ination weakens the structural judgment at the regional level, leading to isolated points and holes in the segmentation results. On the other hand, SAM, which learns in a self-supervised manner, lacks the guidance of dense labels and semantic information, but possesses strong generalization and adaptability, emphasizing the capture of structured information within the data. As a result, its segmentation results exhibit greater coherence and completeness. Based on this, this paper proposes SAMEnhancer, which optimizes segmentation results by combining the strong semantic capability of label-guided network training and the strong structural connectivity of the SAM model. The process is illustrated in Figure 3.

The SAMEEnhancer takes the image I and network segmentation result $\tilde{\mathbf{Y}}$ as input, $\tilde{\mathbf{Y}} \in \mathbb{R}^{C \times H \times W}$, $\tilde{\mathbf{Y}}_{c,h,w} \in [0, 1]$, which represents the confidence score. C represents the number of categories; for road segmentation, it is set to 2. H and W represent the image height and width, respectively. After processing with SAMEEnhancer, the tensor shape remains the same, and the values are binarized to 0 and 1. The SAMEEnhancer operates through a three-stage process, which will be explained in the following sections.

324 3.2.1 POINT PROMPT EXTRACTION
325

326 The first step is to extract points, which serves to provide prompts for SAM in the subsequent steps.
 327 The prompts need to meet the following conditions: 1. Inside the target area, 2. Close to the center,
 328 and 3. With high confidence. Thus, we selected three points as the prompt point set: the centroid
 329 indicates the center of the polygon weighted by confidence; the center of the minimum enclosing
 330 circle indicates the morphological center of the polygon; to avoid the first two points falling outside
 331 the polygon, we add the point of highest confidence inside. The mathematical process is as follows:

332 Specifically, the output image y^{pred} is first subjected to morphological transformations, performing
 333 opening followed by closing operations:

$$335 \quad \text{morph}(\tilde{\mathbf{Y}}) = \text{close}(\text{open}(\tilde{\mathbf{Y}})) \quad (1)$$

337 which can remove isolated points in the prediction results and smooth the boundaries of regions,
 338 thereby avoiding noise from affecting keypoint extraction. The opening and closing operations are
 339 defined by equations 2 and 3, respectively:

$$341 \quad \text{open}(\tilde{\mathbf{Y}}) = (\tilde{\mathbf{Y}} \ominus k^{open}) \oplus k^{open} \quad (2)$$

$$342 \quad \text{close}(\tilde{\mathbf{Y}}) = (\tilde{\mathbf{Y}} \oplus k^{close}) \ominus k^{close} \quad (3)$$

344 where \ominus denotes the erosion operation, \oplus denotes the dilation operation, and k represents the struc-
 345 turing element kernel. Then, find the polygon areas A^P in the transformed prediction mask:

$$347 \quad A^P = \text{findPoly}(\text{morph}(\tilde{\mathbf{Y}})) \implies \{A_1^P, A_2^P, \dots, A_k^P\} \quad (4)$$

349 in which, A^P represents k binarized polygonal regions, and we use the prediction confidence C of
 350 each pixel as the value of the polygon, resulting in:

$$352 \quad V_i^P = C \times A_i^P \quad (5)$$

354 In each polygon P_i , three feature points are identified: the centroid pt^{cent} , the circumcircle
 355 center pt^{exc} , and the highest confidence internal point pt^{hconf} , forming the feature point set:
 356 $P_i = [pt_i^{cent}, pt_i^{exc}, pt_i^{hconf}]$. pt_i^{cent} is defined as equation6.

$$358 \quad 359 \quad pt_i^{cent} = (x_i^{cent}, y_i^{cent}) = \left(\frac{\sum_{x=0}^{W-1} \sum_{y=0}^{H-1} x \cdot V_i^P(x, y)}{\sum_{x=0}^{W-1} \sum_{y=0}^{H-1} V_i^P(x, y)}, \frac{\sum_{x=0}^{W-1} \sum_{y=0}^{H-1} y \cdot V_i^P(x, y)}{\sum_{x=0}^{W-1} \sum_{y=0}^{H-1} V_i^P(x, y)} \right) \quad (6)$$

361 pt_i^{exc} and pt_i^{hconf} are the center of the minimum enclosing circle and the internal point with the
 362 highest confidence, as given in Equation 7 and Equation8, respectively:

$$365 \quad pt_i^{exc} = \text{minEnclosingCircle}(A_i^P) \quad (7)$$

$$368 \quad pt_i^{hconf} = \text{argmax}(V_i^P) \quad (8)$$

369 Finally, remove the points that are not inside the polygon.

371 3.2.2 SEGMENTATION OPTIMIZATION VIA SAM
372

373 In the second stage, we utilize the extracted internal points as inputs to the mobile SAM. SAM is
 374 developed using self-supervised training, leveraging a large dataset of over 1 billion images. This
 375 extensive training allows SAM to achieve excellent generalization across diverse segmentation tasks.
 376 It can effectively utilize various prompts like points and masks, including points and bounding boxes,
 377 enabling users to interactively guide the segmentation process. However, SAM is relatively slow in
 terms of processing speed. Mobile SAM addresses this issue by enhancing speed while maintaining

378 the model’s performance, making it more suitable for real-time applications. In this paper, we aim
 379 to achieve automatic extraction without human interaction by using the key points from the previous
 380 section as prompts input into Mobile SAM, thereby obtaining the generated masks \hat{Y} . The formula
 381 is shown in Equation 9.
 382

$$\hat{Y} = \text{mobile SAM}(I, P) \quad (9)$$

385 3.2.3 RESULT FUSION

387 The final stage of the SAMEnhancer process involves fusing the mobile SAM segmentation results
 388 with the original predictions from the network. This fusion is guided by confidence scores, in \tilde{Y} , the
 389 portions with confidence above the threshold (0.9 in this paper) are retained, while the other parts
 390 keep the results from \tilde{Y} . The fusion process can be mathematically represented as Equation 10.
 391

$$Y^{\text{merge}} = \begin{cases} \tilde{Y} & \text{if } V^P > 0.9 \implies \text{High semantic confidence} \\ \hat{Y} & \text{if } V^P \leq 0.9 \implies \text{Strong structural coherence} \end{cases} \quad (10)$$

395 By integrating the strengths of both the original and refined outputs, the SAMEnhancer generates
 396 a final segmentation map that maintains high accuracy and consistency, particularly in challenging
 397 scenarios with diverse road conditions.

398 In summary, the SAMEnhancer tool is designed to optimize segmentation outcomes by effectively
 399 combining the strengths of initial predictions with the results generated through mobile SAM. This
 400 approach not only enhances segmentation accuracy but also improves the overall performance of
 401 segmentation in unstructured autonomous driving applications.

403 4 EXPERIMENTS AND RESULTS

404 4.1 EXPERIMENT SETTING

407 A series of experiments were conducted in this study to validate the effectiveness of ADC learning
 408 and SAMEnhancer. The datasets used were IDD and Bdd100K. For the object detection task, the
 409 objects were reclassified into five categories: motor vehicles, non-motor vehicles, people, animals,
 410 and traffic signs. For the drivable area, no distinction was made between road and drivable fall-
 411 back. The experiments were based on the PyTorch framework, utilizing torch 2.0.1+cu117 and a
 412 GeForce 4090 GPU. The experimental setup was based on the YOLOP network architecture, with
 413 the lane detection branch removed. The following sections will provide a detailed description of the
 414 validation experiments for ADC learning and SAMEnhancer.

416 4.2 RESULTS

418 Figure 4 presents the visualization results of the proposed method based on YOLOP. The first row
 419 shows the results of training the IDD dataset using only alternate training, the second row presents
 420 the outcomes of training with ADC learning, and the third row presents the results after processing
 421 with SAMEnhancer. It can be observed that the ADC learning strategy reduces the occurrence of
 422 perception errors, while SAMEnhancer extracts a more coherent and complete drivable surface on
 423 this basis.

424 4.3 ADC LEARNING

426 This study first validated that the performance of the model declines when using non-overlapping
 427 annotated data solely with rotation training, as shown in Table 1. The Bdd100K dataset was utilized,
 428 which is a city road dataset suitable for multi-task learning. The complete dataset was used for
 429 experiments and then divided into two parts: one with segmentation labels and the other with de-
 430 tection labels, simulating the scenario of non-overlapping annotations. Subsequently, ADC learning
 431 was applied to this data labeling, resulting in the outcomes shown in Table 1. Recall and mAP are
 evaluation metrics for object detection, while mIoU is an evaluation metric for road segmentation.

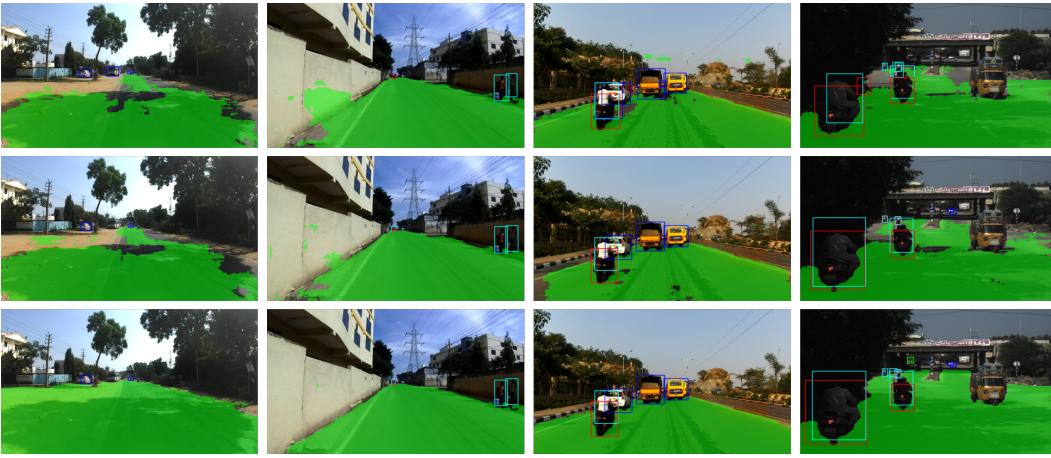


Figure 4: The visualization experiment results. The top row shows the results of YoloP on IDD dataset without ADC learning, the middle row presents the result with ADC learning, and the bottom row shows the results with ADC learning and SAMEnhancer.

It can be observed that the model’s performance decreased under non-overlapping annotations, but improved after employing the proposed ADC learning. A similar phenomenon was also observed in the IDD dataset with the same experiment.

Secondly, this study validated the effectiveness of ADC learning under different network architectures. The encoder of YOLOP was replaced with ConvNeXt, EfficientNet, and DenseNet, respectively. The experimental results are shown in Table 2, demonstrating that ADC learning enhances model performance across various network structures.

Table 1: Ablation study on ADC learning with Bdd100K and IDD dataset. Recall and mAP50 are detection metrics, and mIOU is the segmentation metric.

	Bdd100KYu et al. (2020)			IDDVarma et al. (2019)		
	Recall	mAP50	mIOU	Recall	mAP50	mIOU
All label	89.2	76.5	91.5	-	-	-
Non-overlap label	86.4	73.3	88.8	52.8	25.4	91.3
ADC learning	87.1	74	88.9	55.2	29.3	93.8

Table 2: Ablation study on ADC learning with different encoder architectures.

Encoder	without ADC learning			with ADC learning		
	Recall	mAP50	mIOU	Recall	mAP50	mIOU
ConvNeXtLiu et al. (2022)	52	24.8	88.9	53.8	28.8	91.6
EfficientNetKoonce & Koonce (2021)	51.9	25.4	90.4	53.7	29.1	92.3
DenseNetHuang et al. (2018)	47.6	22	86.2	50.3	26.2	90.5
CSPDarknet(YoloP)Wu et al. (2022)	52.8	25.4	91.3	55.2	29.3	93.8

4.4 SAMENHANCER

This subsection tested the effects of using SAMEnhancer versus not using it. Table 3 presents the experimental results. In the table, ACC refers to accuracy, IoU indicates the performance of individual categories within the drivable area, and mIoU represents the average value between the drivable area and the background. The first row, "Network prediction," lists the metrics for the network’s prediction results, the second row, "Mobile SAM prediction," shows the metric calculations for the

486 output values from the SAMEnhancer sampled points and mobile SAM predictions, and the third
 487 row, "Merge result" represents the final output after combining both results.
 488

490 Table 3: The ablation study on SAMEnhancer. The top row shows the network output results,
 491 the middle row demonstrates the mobile SAM output results, and the bottom row depicts the final
 492 merged results.

	ACC	IOU	mIOU
Network Prediction	96.9	89.5	93.8
Mobile SAM Prediction	96	88	91.2
Merge result	97.3	91.6	94.5

493 5 CONCLUSION

501 This study investigates multi-task learning in unstructured scenes and proposes Anti-Degradation
 502 Complementary learning (ADC learning) and SAMEnhancer. ADC learning addresses the scattered
 503 nature of unstructured scene datasets by designing a two-phase training method with joint non-
 504 overlap annotations, combined with semi-supervised learning. SAMEnhancer enhances the recog-
 505 nition performance of roads in unstructured scenes by leveraging the strong structural understand-
 506 ing of SAM along with the semantic understanding capability of the network itself, thereby addressing
 507 the fragmentation issue of roads in unstructured scenes. Qualitative and quantitative experiments
 508 demonstrate the effectiveness of the proposed methods.

510 REFERENCES

- 511 Bhakti Baheti, Shubham Innani, Suhas Gajre, and Sanjay Talbar. Semantic scene segmentation in
 512 unstructured environment with modified deeplabv3+. *Pattern Recognition Letters*, 138:223–229,
 513 2020.
- 515 Chih-Yung Chang, Hsu-Ruey Chang, Chen-Chi Hsieh, and Chao-Tsun Chang. Ofrd: obstacle-free
 516 robot deployment algorithms for wireless sensor networks. In *2007 IEEE Wireless Communica-
 517 tions and Networking Conference*, pp. 4371–4376. IEEE, 2007.
- 518 Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-
 519 decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of
 520 the European conference on computer vision (ECCV)*, pp. 801–818, 2018.
- 522 Kaiwen Duan, Lingxi Xie, Honggang Qi, Song Bai, Qingming Huang, and Qi Tian. Location-
 523 sensitive visual recognition with cross-iou loss. *arXiv preprint arXiv:2104.04899*, 2021.
- 524 Palmani Duraisamy and Sudha Natarajan. Multi-sensor fusion based off-road drivable region detec-
 525 tion and its ros implementation. In *2023 International Conference on Wireless Communications
 526 Signal Processing and Networking (WiSPNET)*, pp. 1–5. IEEE, 2023.
- 528 Gao Huang, Shichen Liu, Laurens Van der Maaten, and Kilian Q Weinberger. Condensenet: An
 529 efficient densenet using learned group convolutions. In *Proceedings of the IEEE conference on
 530 computer vision and pattern recognition*, pp. 2752–2761, 2018.
- 531 Zhijian Huang, Sihao Lin, Guiyu Liu, Mukun Luo, Chaoqiang Ye, Hang Xu, Xiaojun Chang, and
 532 Xiaodan Liang. Fuller: Unified multi-modality multi-task 3d perception via multi-level gradient
 533 calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.
 534 3502–3511, 2023.
- 535 Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete
 536 Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceed-
 537 ings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.
- 539 Brett Koonce and Brett Koonce. Efficientnet. *Convolutional neural networks with swift for Tensor-
 flow: image recognition and dataset categorization*, pp. 109–123, 2021.

- 540 Der-Hau Lee and Jinn-Liang Liu. End-to-end multi-task deep learning and model based control
 541 algorithm for autonomous driving. *arXiv preprint arXiv:2112.08967*, 2021.
 542
- 543 Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie.
 544 A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and*
 545 *pattern recognition*, pp. 11976–11986, 2022.
- 546 John Phillips, Julieta Martinez, Ioan Andrei Bârsan, Sergio Casas, Abbas Sadat, and Raquel Urtasun.
 547 Deep multi-task learning for joint localization, perception, and prediction. In *Proceedings of the*
 548 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4679–4689, 2021.
- 549 Yeqiang Qian, John M Dolan, and Ming Yang. Dlt-net: Joint detection of drivable areas, lane lines,
 550 and traffic objects. *IEEE Transactions on Intelligent Transportation Systems*, 21(11):4670–4679,
 551 2019.
- 552 Jahangir Hossain Setu, Mahmudul Islam, Syed Tangim Pasha, Nabarun Halder, Ekram Hossain,
 553 Asif Mahmud, Ashraful Islam, Md Zahangir Alam, and M Ashraful Amin. Segment anything
 554 model (sam 2) unveiled: Functionality, applications, and practical implementation across multiple
 555 domains. 2024.
- 556 Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multi-
 557 net: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE intelligent vehicles*
 558 *symposium (IV)*, pp. 1013–1020. IEEE, 2018.
- 559 Abhinav Valada, Gabriel L Oliveira, Thomas Brox, and Wolfram Burgard. Deep multispectral se-
 560 mantic scene understanding of forested environments using multimodal fusion. In *2016 Interna-*
 561 *tional Symposium on Experimental Robotics*, pp. 465–477. Springer, 2017.
- 562 Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawa-
 563 har. Idd: A dataset for exploring problems of autonomous navigation in unconstrained envi-
 564 ronments. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp.
 565 1743–1751. IEEE, 2019.
- 566 Maggie Wigness, Sungmin Eum, John G Rogers, David Han, and Heesung Kwon. A rugd dataset
 567 for autonomous navigation and visual perception in unstructured outdoor environments. In *2019*
 568 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5000–5007.
 569 IEEE, 2019.
- 570 Dong Wu, Man-Wen Liao, Wei-Tian Zhang, Xing-Gang Wang, Xiang Bai, Wen-Qing Cheng, and
 571 Wen-Yu Liu. Yolop: You only look once for panoptic driving perception. *Machine Intelligence*
 572 *Research*, 19(6):550–562, 2022.
- 573 Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madha-
 574 van, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learn-
 575 ing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp.
 576 2636–2645, 2020.
- 577 Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and
 578 Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications.
 579 *arXiv preprint arXiv:2306.14289*, 2023a.
- 580 Chunhui Zhang, Li Liu, Yawen Cui, Guanjie Huang, Weilin Lin, Yiqian Yang, and Yuehong Hu.
 581 A comprehensive survey on segment anything model for vision and beyond. *arXiv preprint*
 582 *arXiv:2305.08196*, 2023b.
- 583 Zhuoyang Zhang, Han Cai, and Song Han. Efficientvit-sam: Accelerated segment anything model
 584 without performance loss. *arXiv preprint arXiv:2402.05008*, 2024.
- 585 Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang.
 586 Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023.
- 587