SEMI-SUPERVISED SEMANTIC SEGMENTATION ON VEHICLE-MOUNTED FISH-EYE CAMERA IMAGES

Sneha Paul, Corresponding Author
Ph.D. Student, Concordia Institute for Information Systems Engineering
Computer Science and Visual Arts Integrated Complex, 1515 St. Catherine W., Montreal, Canada
Email: sneha.paul@mail.concordia.ca

Zachary Patterson
Professor, Concordia Institute for Information Systems Engineering
Computer Science and Visual Arts Integrated Complex, 1515 St. Catherine W., Montreal, Canada
Tel: 514-848-2424 ext. 8830, Email: zachary.patterson@concordia.ca

Nizar Bouguila
Professor, Concordia Institute for Information Systems Engineering
Computer Science and Visual Arts Integrated Complex, 1515 St. Catherine W., Montreal, Canada
Tel: 514-48-2424 ext. 5663, Email: nizar.bouguila@concordia.ca

Word Count: 4877 words + 7 table(s) $\times$ 250 = 6627 words

Submission Date: May 14, 2026

ABSTRACT

The use of large Field of View (FoV) cameras employing fish-eye lenses benefits various computer vision applications in autonomous driving. Although deep learning has been successful in various computer vision applications using regular perspective images, its potentiality in fish-eye camera applications still needs to be explored due to the availability of relatively small datasets for fully-supervised learning. Semi-supervised learning can be a potential solution to address this challenge. In this study, we focus on exploring the use of semi-supervised learning for the semantic segmentation of fish-eye images. To the best of our knowledge, this is the first work on semi-supervised semantic segmentation of fish-eye images. We implement three popular semi-supervised methods from the perspective image literature for fish-eye image segmentation and conduct extensive analysis in different settings. Based on the best practices observed in these methods, we propose three variants of semi-supervised segmentation models for fish-eye images. Our proposed approaches are evaluated on the WoodScape dataset, which is collected from vehicle-mounted fish-eye cameras. Detailed experiments demonstrate an 8.89% enhancement of the proposed semi-supervised method compared to fully supervised learning using the same amount of labeled data. We also perform an extensive sensitivity analysis to exhibit the effectiveness and advantages of the proposed method in this study.

Keywords: Semi-supervised learning, Semantic Segmentation, fish-eye images, Autonomous driving, Autonomous vehicles, AED50 - Committee on Artificial Intelligence and Advanced Computing.

INTRODUCTION

Currently, various computer vision applications in autonomous driving are benefiting from the use of large Field of View (FoV) cameras (*1*). These types of cameras use fish-eye lenses, also known as wide-angle lenses, to capture a broad view by using extensive non-linear mapping instead of regular perspective projection. However, this results in a strong radial distortion to the captured images. One potential solution to the distortion issue is to use transformation techniques to convert fish-eye images into regular perspective images (*2*). However, this compromises the rich feature information contained at the boundary of the images. As a result, the detection process is hindered. Another possible solution is to use the distorted fish-eye images directly in the model, as suggested by some existing research (*3, 4*). However, the current literature on computer vision is not well-optimized for learning from fish-eye cameras, thus providing a need for further research in this area.

Semantic segmentation, an important computer vision task, focuses on classifying each pixel in an image and grouping together the pixels that belong to a specific object. Thus it segments a scene into bounding areas or masks containing different objects. 2D perspective images are the very first domain to which semantic segmentation was used in the context of autonomous driving (*5*). However, the application of segmentation in autonomous driving has been extended to other domains as well, e.g. panoptic segmentation (*6*), point cloud (*7*), depth images (*8*), thermal images (*9*), and so on. Various segmentation datasets on autonomous driving, e.g. Cityscape (*10*), KITTI (*11*), Toronto City (*12*), ApolloScape (*13*), WoodScape (*1*) have leveraged the computer vision research on autonomous driving. The significance of semantic segmentation in autonomous driving lies in its ability to enable the vehicle's sensor systems to precisely comprehend and differentiate various objects and areas in the surroundings. It leads to well-informed decisions based on specific information about the road, obstacles, pedestrians, traffic signs, and other crucial elements, ultimately ensuring safer and more authentic autonomous navigation.

However, the application of semantic segmentation for autonomous driving on fish-eye images is not as well-explored as in other domains. Very little research on fish-eye image segmentation is available in the literature, which has previously explored different approaches to address the problem of radial distortion in fish-eye images (*14–16*) due to non-linear mapping. Most of these researches rely on models trained on synthetic fish-eye images, which are obtained by applying different fish-eye transformations to the perspective images. However, these approaches do not perform well when using real-world fish-eye data. Therefore, this research focuses on learning semantic segmentation from real-world fish-eye data. Moreover, existing literature on semantic segmentation on fish-eye autonomous driving images focuses largely on multi-modal, multi-task models (*1, 3, 17*), gaining an advantage in learning pixel-level representation from different modalities and similar tasks. However, these models are computationally expensive due to the high number of learnable parameters. To this end, a single-task, uni-modal model will be a potential solution. In addition, the current literature on semantic segmentation focuses on supervised learning (*2, 15, 18*), which requires a large amount of labeled data to train deep learning models effectively.

Since semantic segmentation involves classifying each pixel in an image, annotation is more costly and laborious than other computer vision tasks, such as object classification. Consequently, there is a lack of publicly available semantically annotated fish-eye segmentation datasets. Semi-supervised learning has emerged as a potential solution to this problem, allowing models to learn from a small amount of labeled data while leveraging a larger amount of unlabeled data.

However, to the best of our knowledge, there is no existing work on semi-supervised semantic segmentation for fish-eye data. Therefore, in our study, we explore three existing methods for semi-supervised segmentation (*19–21*) on fish-eye images and combine the best practices to propose a new model for fish-eye semantic segmentation. We also perform an in-depth sensitivity analysis on different model-specific hyperparameters of our proposed methods. We evaluate our proposed method on the WoodScape dataset (*1*), which consists of real-world fish-eye images. We show an 8.89% improvement over fully supervised learning with the same amount of labeled data by applying the proposed method.

The key contribution of this work can be summarized as follows:

- We present the first study on semi-supervised segmentation in the context of fish-eye camera images in the context of autonomous vehicles or otherwise.
- We adopt different popular semi-supervised segmentation methods for this task and compare and benchmark their performance.
- We also present a detailed sensitivity study on each of the methods in this study for further understanding of the impact of different components.
- From the study on existing methods, we combine the best practices and propose a new method for semi-supervised fish-eye camera segmentation that outperforms all off-the-shelf methods for this task.

RELATED WORKS

In this section, we will discuss the existing literature on semantic segmentation models for autonomous driving. Then we will elaborate on the available work in semantic segmentation on regular perspective and fish-eye images in other domains that are relevant to our work.

Segmentation for Autonomous Driving

Various existing work deals with semantic segmentation for autonomous driving scenarios from supervised settings. For example, (*22*) and (*23*) worked on multi-class semantic segmentation, whereas (*24*) focused on road segmentation. (*25*) proposed a fully-convolutional network (FCN) based architecture, a one-stage pipeline using fully-connected layers to replace the convolutional layers for predicting coarse output maps. These coarse maps are upsampled by deconvolution to produce dense pixel-level labels. (*23*) extended this idea by using a encoder-decoder-based CNN network. However, these networks work in supervised settings, where labeled data is necessary, making semantic segmentation a laborious and costly task.

To this end, a great deal of research on unsupervised settings has been proposed, requiring no (*26*), or little (*27*) labeled data. The Asynchronous Teacher-Student Optimization (ATSO) algorithm (*28*) used continual learning by partitioning unlabeled training data into two subsets. One subset of the training data was used for fine-tuning the model that updated the labels of the other subset. (*29*) proposed a semi-supervised segmentation network using adversarial learning. (*30*) proposed a multi-task, self-supervised segmentation network that is efficient for classification, segmentation, and detection tasks when labeled data is unavailable. Though these models perform satisfactorily on perspective images, their performance deteriorates for fish-eye images due to high radial distortion. Thus in-depth study is necessary on the unsupervised segmentation domain for fish-eye autonomous driving images.

Perspective Image Segmentation

Semantic segmentation is a popular dense prediction task in computer vision that has seen a significant number of developments in the past few years, including CNN and Vision transformer (ViTs) based models. Vision Transformers are gaining popularity in Semantic segmentation for the advantage gained from their special network architecture. These models learn pixel-level feature representation using a transformer-based encoder along with an attention mechanism and a bottleneck layer (*31, 32*). SegViT (*33*) stands apart from conventional ViTs due to its use of an Attention-to-Mask (ATM) module. This module effectively transfers learned similarity maps, derived from trainable parameters and spatial feature maps, into segmentation masks. They also propose a shrunk structure of the SegViT model based on a query-based up-sampling (QU) and query-based down-sampling (QD) module that reduces a model's computational cost by 40%. SegFormer (*34*) is another ViT-based semantic segmentation model that incorporates a transformer-based encoder with a lightweight full MLP-based decoder. SegFormer does not require positional encoding that needs to be interpolated, resulting in reduced performance of the model when the resolution of training and testing data is different. Thus it introduces Mix-FFN, which is a combination of convolution block with MLP that can still capture locational information without positional encoding. None of these SOTA models have been explored in the concept of fish-eye image segmentation.

Fish-eye Image Segmentaiton

There has been a bit of research on fish-eye image segmentation in recent years that addresses distortion at the edge of the image. One such approach, proposed in (*14*), uses an Efficient Residual Factorized Network- (ERFNet) based architecture to handle distortion in synthetic fish-eye data. Another approach, proposed in (*15*), uses an adaptive deformable CNN-based architecture that can adapt to fish-eye images while being trained on pinhole camera images, thus avoiding the laborious task of data transformation. However, due to the lack of large public fish-eye datasets, most existing literature converts regular perspective images into fish-eye images. In (*16*), a method is proposed to reuse rectilinear images as fish-eye images to address this issue. In (*18*), a seven-degree of freedom (DoF) augmentation method is proposed, which uses rectilinear images during training to transform them into fish-eye images. This method simulates fish-eye images from rectilinear images captured from different cameras with varying focal lengths, positions, and orientations. Although the seven DoF augmentation method is effective in increasing the model's mean Intersection over Union (mIoU) compared to highly distorted real-world fish-eye data, it falls short in situations where the distortion is not similar among images, which is a common criterion in real-world scenarios. As a result, existing models trained on transformed or synthetic fish-eye data cannot achieve superior performance when using real-world fish-eye data. Another key observation we find from very recent literature is that multi-modal or multi-task data is jointly used to learn based on fish-eye images (*3, 35*). However, such a dataset is costly to collect and computationally demanding to train. Therefore, the focus of our research is to learn semantic segmentation from real-world fish-eye data using the image alone.

In summary, the current research on semantic segmentation in autonomous driving has primarily focused on different modalities, excluding fish-eye images, particularly in situations where labeled data is scarce and annotation is expensive and time-consuming. Moreover, there is a noticeable gap in studies conducted on real-world fish-eye images captured from autonomous vehicles (AVs). To this end, our work introduces a semi-supervised semantic segmentation approach specifically designed for real-world fish-eye images within the domain of autonomous driving.

METHOD

In this work, we explore three well-known semi-supervised segmentation methods that were originally developed for regular images: CPS (*19*), MeanTeacher (*20*), and CPS with CutMix (*36*). An overview of these methods is presented in Figure 1. However, none of these methods are specifically tailored for fish-eye image segmentation. To address this limitation, we propose a novel semi-supervised method that draws inspiration from existing methods in the literature and incorporates best practices discovered during our preliminary studies. Here, we will first provide an overview of existing semi-supervised segmentation methods on perspective images that are used in this study. Next, we will discuss the proposed method.

Overview of Existing Semi-supervised Segmentation Methods

In (*19*), a new approach for semi-supervised learning of semantic segmentation was proposed that learns from the labeled and unlabeled perspective images simultaneously. The approach is based on consistency regularization and is called Cross Pseudo Supervision (CPS). CPS aims to increase the consistency between two different segmentation networks for similar input images. Both labeled and unlabeled data are fed into the two different networks, and for each unlabeled image, the model generates a one-hot pseudo-prediction map. This pseudo-prediction map is then used to supervise the other segmentation network and vice versa. This way, the pseudo-segmentation map is used as an additional signal to provide supervision for the unlabeled data.

Formally, CPS generates the prediction as follows:

$$\mathsf{P}_1 = f(\mathsf{X};\theta_1), \tag{1}$$

$$\mathsf{P}_2 = f(\mathsf{X};\theta_2). \tag{2}$$

Here, P_1 and P_2 are the segmentation prediction obtained from the two segmentation networks after applying the softmax function. Note that the two segmentation networks are similar in structure but initialized and updated differently. The predictions are then used to generate the pseudo-labels. The whole process can be described as follow:

$$\begin{aligned}\mathsf{X} \rightarrow \mathsf{X} \rightarrow f(\theta_1) \rightarrow \mathsf{P}_1 \rightarrow \mathsf{Y}_1 \\ \searrow f(\theta_2) \rightarrow \mathsf{P}_2 \rightarrow \mathsf{Y}_2,\end{aligned} \tag{3}$$

where, Y_1 and Y_2 are one-hot pseudo-labels.

The CPS method imposes two loss functions, a supervised loss $\mathscr{L}_s$ and pseudo supervised loss (unlabeled loss) $\mathscr{L}_{cps}$. The supervised loss, $\mathscr{L}_s$, is a standard cross-entropy loss applied at pixel level.

$$\begin{aligned}\mathscr{L}_s = \frac{1}{|\mathscr{D}^l|}\sum_{\mathsf{X}\in\mathscr{D}^l}\frac{1}{W\times H}\sum_{i=0}^{W\times H}(\ell_{ce}(\mathbf{p}_{1i},\mathbf{y}^*_{1i}) \\ +\ell_{ce}(\mathbf{p}_{2i},\mathbf{y}^*_{2i})),\end{aligned} \tag{4}$$

The pseudo supervised loss (unsupervised loss), $\mathscr{L}_{cps}$ is also a standard cross entropy loss applied bi-directionally on the unlabeled data $\mathscr{D}_u$ and predicts pseudo-labels. The CPS loss can be represented as:

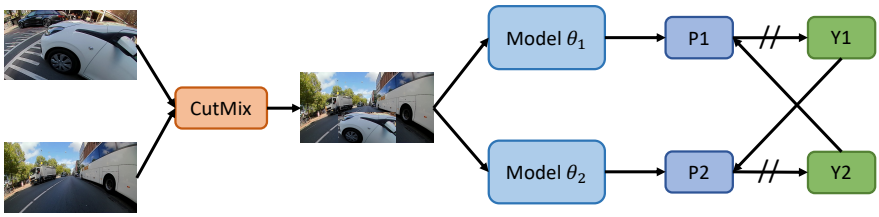

Figure 1 Overview of the semi-supervised segmentation model

$$\mathscr{L}^u_{cps} = \frac{1}{|\mathscr{D}^u|} \sum_{\mathsf{X} \in \mathscr{D}^u} \frac{1}{W \times H} \sum_{i=0}^{W \times H} (\ell_{ce}(\mathbf{p}_{1i}, \mathbf{y}_{2i}) + \ell_{ce}(\mathbf{p}_{2i}, \mathbf{y}_{1i})). \tag{5}$$

The overall cross pseudo supervised loss is the sum of the supervised loss $\mathscr{L}_s$ on the labeled data, $\mathscr{D}_l$ and the pseudo supervised loss $\mathscr{L}_{cps}$ on the unlabeled data $\mathscr{D}_u$.

$$\mathscr{L} = \mathscr{L}_s + \lambda \mathscr{L}_{cps}, \tag{6}$$

where, λ is a hyper-parameter that trade-off the importance of the two losses.

MeanTeacher (*20*) is another semi-supervised learning method that does not train both encoders, which is expensive and less intuitive. Rather, inspired by the semi-supervised literature of object recognition (*37*), it updates the teacher encoder as an exponential moving average (EMA) of the online encoder (student encoder), denoted as $\bar{\theta}$. The loss function of MeanTeacher can be represented as:

$$\begin{aligned} &\mathsf{X} \rightarrow \mathsf{X}_1 \rightarrow f(\theta) \rightarrow \mathsf{P}_1 \\ &\quad \searrow \mathsf{X}_2 \rightarrow f(\bar{\theta}) \nrightarrow \mathsf{P}_2 \end{aligned} \tag{7}$$

where, X_1 and X_2 are two different augmentations of images an input image X.The consistency regularization works in a way to align the consistency map P_1 predicted by the student network to the consistency map P_2 of the teacher network. At training time, the P_1 is supervised by P_2. However, the teacher network does not back-propagate the loss but is updated as the exponential moving average (EMA) of the student network. As a regularization method, EMA helps to stabilize the student model's updates, smoothing out fluctuations and facilitating knowledge transfer to the student network. This technique enhances the model's capacity to handle unlabeled data and improves overall performance on both labeled and unlabeled datasets. $f(\theta)$. It reduces the redundant calculation and cost of training while keeping the weights of the teacher network similar to the student network.

Another semi-supervised method combines CPS with the CutMix augmentation (*36*). CutMix augmented images are fed as input in both of the networks $f(\theta_1)$ and $f(\theta_2)$. The rest of the method is similar to (*37*). Finally, each generated pseudo-label is used to supervise the other segmentation network.

Proposed Method

In this paper, we propose a few variants of the existing methods that are inspired by the existing literature and aim to improve semi-supervised segmentation on fish-eye images for autonomous vehicles. These methods are described in the following subsections. We also attempt to improve the existing methods by tuning important parameters of individual methods for fish-eye images.

CPS with Confidence Threshold

The first variant we propose is inspired by FixMatch (*38*), which involves using confidence thresholding to improve the quality of predicted pseudo-labels on unlabeled samples. FixMatch uses a weakly-augmented image as input to the model, which generates the predicted pseudo-labels. We transform this prediction into a one-hot pseudo-label. The same image is then subjected to strong augmentation and fed into the same model. The one-hot pseudo-label from the weakly-augmented image is used to supervise the prediction of the strongly-augmented image, and this process works as a consistency regularization by ignoring low-confidence pseudo-labels for supervision. The FixMatch objective can be represented as follows:

$$\begin{aligned} &\mathsf{X} \rightarrow \mathsf{X}_s \rightarrow f(\theta) \rightarrow \mathsf{P}_s \nwarrow \\ &\quad \searrow \mathsf{X}_w \rightarrow f(\theta) \rightarrow \mathsf{P}_w \nrightarrow \mathsf{Y}_w, \end{aligned} \tag{8}$$

where, X_w and X_s are the weak and strong augmentations of image X.

We then combine the pseudo-label thresholding idea with CPS. In CPS, one encoder generates pseudo-labels that are used as ground truth for the other branch. By incorporating the thresholding concept of FixMatch (*38*), we only consider pseudo-labels that exceed a pre-defined threshold. Thus providing more confidence in the model's performance.

CPS with CutMix and Confidence Threshold

As mentioned above, using CutMix augmentation with CPS improves the performance of CPS by making the dataset more challenging for the model to learn representation. So, we propose a new variant by introducing Cutmix augmentation into CPS with Confidence Thresholding.

CPS with CutMix, Confidence Threshold, and Hard Augmentations

In this variant, we introduce the concept of hard augmentations with the previously proposed variant. Hard argumentation regularizes a model to prevent overfitting on small labeled samples. To this end, we combine the following augmentations to form a hard augmentation module: RandomCrop; ColorJitter with brightness, contrast, saturation, hue, and Gaussian blur. The model applies Cutmix and hard augmentations on the input and applies threshold on the predicted pseudo-labels.

Tunning Existing Methods

From our preliminary studies, we identify a few key parameters that are significant to the existing methods. Since the existing methods are trained and optimized for perspective images, they do not perform well on fish-eye images. To this end, we propose to change some of these model-specific parameters to tune them for the fish-eye segmentation task. For this study, we focus on parameters of individual methods that we find important for fish-eye segmentation. CPS, with the Cutmix augmentation method, contains two important model-specific parameters: the number of cutmixed image boxes and the range of these boxes. In our study, we tune both of the parameters to perform better in the fish-eye image segmentation. Similarly, for CPS, we tune the CPS weight and the learning rate. For MeanTeacher, we tune the momentum and unsupervised loss weight. For the other three proposed methods, e.g. CPS with confidence thresholding, CPS with cutmix, confidence thresholding, and CPS with cutmix, confidence thresholding, and hard augmentation, the threshold value is a crucial hyperparameter that represents a pre-defined threshold to which only the exceeded pseudo labels are considered.

Dataset Description

WoodScape (*1*) is a dataset specifically designed for autonomous driving tasks captured using multi-camera fish-eye lenses. It comes with extensive ground-truth annotations for nine essential tasks, including semantic segmentation, which covers classes like roads, pedestrians, and traffic signs. The main purpose behind WoodScape is to offer a comprehensive platform for evaluating computer vision algorithms on fish-eye images. The dataset comprises four surround-view cameras and nine different tasks, such as depth estimation, 3D bounding box detection, and soiling detection. Unlike relying on rectification methods, WoodScape encourages the adoption of computer vision techniques tailored explicitly for fish-eye cameras.

This dataset stands out as the first labeled dataset specifically designed for semantic segmentation. The data was collected from three distinct geographical regions: the US, Europe, and China. To ensure diverse sensor mechanical configurations, the majority of the data was gathered from saloon vehicles, while a significant subset comes from sports utility vehicles. The driving scenarios encompass various use cases, including highway driving, urban environments, and parking scenarios. The dataset provides both intrinsic and extrinsic calibrations for all sensors, along with timestamp files for data synchronization. Quality checks were conducted at every stage of data collection, and reviewers performed quality assurance for the annotation data. The sensors used in recording this dataset include four 1MPx RGB fish-eye cameras with a wide 190° horizontal field of view each, one LiDAR rotating at 20Hz (Velodyne HDL-64E), one GNSS/IMU setup (NovAtel Propak6 and SPAN-IGM-A1), one GNSS Positioning with SPS (Garmin 18x), and Odometry signals from the vehicle bus. Overall, WoodScape promises high-quality data and opens avenues for developing unified multi-task and multi-camera models for advancing autonomous driving technology.

Implementation Details

We follow OmniDet and SemiSeg for the training setup and default hyper-parameters. We adopt the Omnidet (*35*) model as a backbone encoder-decoder for the segmentation model. The visual encoder backbone of OmniDet is a ResNet50 network. For semi-supervised training, we first divide the total training samples from WoodScape (*1*) into supervised and unsupervised splits. More specifically, we used 20% of the 8,029 samples as our labeled set and the rest of the samples as unlabeled set. We train the model for 125 epochs with an Adam optimizer and a learning rate of 0.0001, and a batch size of 16. A multi-step scheduler with a decay factor of 0.1 and a step size of [100, 110] is used to adjust the learning rate. The model is trained on an NVIDIA RTX A6000 GPU.

RESULTS

This section presents the results of this study. First, we will discuss the performance of existing methods in fish-eye image segmentation, which we consider the baseline of the study. Then we will present a discussion of the results of our proposed methods.

Baseline

For this and all subsequent experiments, we consider supervised learning with an equal amount of labeled data (20% to total data) used in the semi-supervised methods as the baseline. So, the improvement with the semi-supervised method is purely from using unlabeled data (as the number of labeled samples is the same in both settings). The first row of Table 1 shows the baseline

Table 1 Comparison between fully-supervised learning and semi-supervised learning with unconstrained unlabeled data.

Method	mIoU
Fully-supervised (baseline)	54.32
CPS (from SemiSeg (*19*))	60.31
MeanTeacher (*20*)	54.20
CPS+CutMix (*36*)	62.47
Ours (CPS+Conf. thr.)	60.66
Ours (CPS+CutMix+Conf. thr.)	62.38
Ours (CPS+Conf. thr.+Hard. Aug.)	59.43
Ours CPS+CutMix (optimized)	**63.21**

mIoU for our work, which is 54.32. For the rest of the paper, we will discuss the improvement of different semi-supervised methods over this baseline.

Main Results

Table 1 summarizes the results for all the existing methods as well as our proposed variants. For evaluation, we follow the protocol described in (*19*). More precisely, the evaluation process is unimodal; single-scale testing. The results are presented using the mean intersection-over-union (mIoU) metric, which evaluates model accuracy by computing the intersection-to-union ratio between predicted and ground truth masks for each class. The mIoU offers an average assessment of segmentation performance across all classes, with superior values denoting higher accuracy. As mentioned in the Method section, we re-implement three popular semi-supervised segmentation methods from the regular image literature; namely, CPS (*19*), MeanTeacher (*20*), CPS with CutMix (*36*). From the existing methods, we find the best result for CPS+CutMix. It obtains an accuracy of 62.47, which is an 8.15 improvement of the supervised learning with a sample amount of labeled samples. The next best performance is shown by the original CPS with a mIoU of 60.31. However, MeanTeacher does not perform well, and the result is worse than learning from supervised loss only. Based on these three semi-supervised segmentation methods, we propose three new variants, namely CPS+Confidence Thresholding, CPS+Cutmix+Confidence Thresholding, and CPS+Cutmix+Confidence Thresholding+Hard Augmentation. Among these three proposed methods, the highest performance is obtained from the CPS+Cutmix+Confidence Thresholding variant, which is 62.38%. We do not observe comparable results for the CPS+Cutmix+Confidence Thresholding+Hard Augmentation method. We suspect that imposing hard augmentation to the dataset alters the data distribution and makes it challenging for the model to learn representation. However, the highest mIoU of the model in this study, which is **63.21%**, is achieved by our optimized CPS+Cutmix method.

Table 2 Performance of CPS with Confidence Thresholding for different threshold values

Method setting	mIoU
threshold= 0.5	60.56
threshold= 0.75	60.66
threshold= 0.9	59.96

Table 3 Performance of CPS with CutMix and Confidence Thresholding for different threshold values

Method setting	mIoU
threshold= 0.5	62.38
threshold= 0.75	61.98
threshold= 0.9	62.12
threshold= 0.95	55.60

Analysis of the Proposed Method

In this section, we report the performance of our proposed semi-supervised learning methods on the WoodScape dataset.

CPS with Confidence Thersholding

Table 2 shows the performance of CPS with Confidence Thresholding for different threshold values. The results indicate that the highest mIoU value of 60.66 is obtained with a threshold of 0.75, whereas the lowest value of 59.96 is obtained with a threshold of 0.9. Since the threshold value works as a consistency regularizer by ignoring low-confidence pseudo-labels, the higher threshold value induces a higher regularization effect on the model and thus affects the model's performance by ignoring a larger amount of predicted pseudo-labels for supervision.

CPS with CutMix and Confidence Thersholding

Table 3 demonstrates the performance of CPS with CutMix and Confidence Thresholding at various threshold values. We consider the threshold values of 0.5, 0.75, 0.9, and 0.95 in this study. The results indicate that the highest mIoU is achieved with a threshold of 0.5, while the lowest was observed with a threshold of 0.95. We believe the reason behind the following performance of the model is similar to that explained in the previous subsection.

CPS with CutMix, Confidence Thresholding, and Hard Augmentations

Table 4 represents the performance of our proposed final variant, CPS with CutMix, Confidence Thresholding, and Hard Augmentations. The table shows the mIoU scores for the method with different thresholds: 0.5, 0.75, 0.9, and 0.95. Here, we see an interesting observation as the model's performance becomes random with the different threshold values. We suspect that the

Table 4 Performance of CPS with CutMix, Confidence Thresholding, and Hard Augmentations for different threshold values

Method setting	mIoU
threshold= 0.5	58.37
threshold= 0.75	57.41
threshold= 0.9	59.43
threshold= 0.95	46.01

Table 5 Comparison among various combinations of CPS weight and learning rate values for CPS method.

	CPS Weight		
LR	1.5	1	0.5
0.01	47.36	46.46	46.67
0.001	57.28	57.19	56.10
0.0001	60.31	59.86	60.40

application of hard augmentation has made the dataset challenging enough for the model to learn.

Sensitivity Study on Tunning Existing Method

In this section, we discuss the results of the sensitivity study on existing methods that we find important for the performance in fish-eye segmentation.

Table 5 shows different combinations of CPS weight and learning rate values for the CPS method. The table includes the performance of the CPS method of different combinations of CPS weights of 1.5, 1, and 0.5. We can see that with the change in the CPS weights with a fixed learning rate, the model's performance does not change significantly. However, by decreasing the learning rates from 0.01 to 0.0001 with a constant CPS weight, the model's performance increases significantly. The highest mIoU, 60.40% is achieved with a 0.0001 learning rate with 0.5 CPS weight. So, it can be concluded that the model's performance is more sensitive to the learning rate compared to CPS weight.

In Table 6, we can see the performance of EMA as a teacher when the unsupervised weight and momentum values were adjusted. The table includes the method settings and their corresponding mIoU scores. The default setting, with W=100 and m=0.99, scored 54.20. We also tested other settings with different values of W and m, resulting in different scores. From the analysis, it can be seen that as the unsupervised weights (W) get lower, the model shows better performance. The highest accuracy in this setting, 57.15%, is achieved from an unsupervised weight of 1.0. For the momentum values, with 0.999 momentum, the model achieves better performance, 55.62%, compared with other momentum values. Since momentum adjusts the size of the next update introducing a penalty term with a higher or lower momentum value affects the model's performance.

Table 6 Performance of EMA as a teacher for different values of unsupervised weight and momentum.

Method setting	mIoU
Default (W=100 & m=0.99)	54.20
W=10.0	56.38
W=1.0	57.15
m=0.999	55.62
m=0.9999	46.21

Table 7 Performance of CPS with CutMix augmentation with different combinations of the number of cutmixed boxes and the range of boxes of CutMix augmentation

Method setting	mIoU
Default (box number=3; range= 0.25-0.5)	62.47
nb. of. box=2 and range=0.1 to 0.25	63.13
nb. of. box=2 and range=0.25 to 0.5	62.60
nb. of. box=2 and range=0.4 to 0.75	62.29
nb. of. box=3 and range=0.1 to 0.25	62.02
nb. of. box=3 and range=0.4 to 0.75	62.59
nb. of. box=4 and range=0.1 to 0.25	62.00
nb. of. box=4 and range=0.25 to 0.5	61.91
nb. of. box=4 and range=0.4 to 0.75	**63.21**

To understand the performance of CPS with CutMix augmentation, in table 7, we conduct experiments using different combinations of the number of cutmixed boxes and the range of these boxes. The table demonstrates the method setting and the corresponding mIoU value. This experiment aims to determine the optimal combination of box number and range that would yield the highest mIoU value. The default setting for the experiment had a box number of 3 and a range of 0.25-0.5, with a mIoU value of 62.47. The highest performance, **63.21%**, is achieved by 4 cut mix boxes with a range of 0.4 to 0.75. Also, note that, with the increment of the cutmix range, the performance of the model increases as the overlapping area of the two images increases. Thus the model gets a wider view of the images to capture the underlying information. The highest mIoU of the proposed methods, **63.21%**, is achieved from 4 boxes with a range of 0.4 to 0.75, which is a **8.89%** improvement over the supervised baseline.

Overall, the results of the experiment are promising, as they showed that CPS with CutMix augmentation can significantly improve the performance of the semi-supervised segmentation model for fish-eye images.

(a) Example 1 **(b) Example 2** **(c) Example 3**

Figure 2 Example of segmentation generated by semi-supervised segmentation model.

Qualitative Analysis

Figure 2 show some images generated by our proposed semi-supervised segmentation method (CPS+cutmix optimized). The sample images show perfect detection of most of the objects and their areas.

CONCLUSION

Fish-eye camera images differ from perspective camera images in that they capture a Field of View (FOV) of 180° to 360°. Because of this characteristic, fish-eye images are potentially important for use in autonomous driving. However, the larger FOV comes at a price of higher radial distortion and requires specialized models to handle this challenge. Despite the high potentiality of this area of fish-eye images, it is not well-explored due to the lack of large publicly annotated datasets. Semantic segmentation, which classifies an image at the pixel level, is a popular domain of computer vision. However, data annotation for semantic segmentation tasks is more expensive and laborious compared to object classification. Semi-supervised learning provides a solution to this problem, but it has not been studied in the context of semi-supervised segmentation from fish-eye cameras. In this paper, we proposed a semi-supervised semantic segmentation method for fish-eye images. We explored various existing semi-supervised methods and benchmarked their performances. Based on our findings and best practices from these methods, we proposed three new variants for semi-supervised segmentation. An analysis of these methods shows that semi-supervised methods can improve up to 8.89% over fully-supervised methods for the same amount of labeled data. However, one of the limitations of our work is that we have focused on a subset of the available literature concerning semi-supervised semantic segmentation methods designed for perspective images. As a result, we recognize the potential for enhancing the performance (mIoU) in semantic segmentation tasks for fish-eye images used in autonomous driving by delving into other semi-supervised methods already established in the field. We hope that this work will encourage the development of more specialized techniques for learning semantic segmentation and other computer vision tasks related to fish-eye images, especially when labeled data is scarce in the context of autonomous driving.

ACKNOWLEDGEMENTS

We would like to thank Mitacs Accelerate and BusPas Inc. for funding this research.

AUTHOR CONTRIBUTION STATEMENT

The authors acknowledge their involvement in the paper in the following capacities: Sneha Paul, Zachary Patterson, and Nizar Bouguila contributed to the study's conception and design. Sneha Paul, Zachary Patterson, and Nizar Bouguila conducted the analysis and interpretation of the results. All authors thoroughly reviewed the results and granted approval for the final version of the manuscript.

REFERENCES

1. Yogamani, S., C. Hughes, J. Horgan, G. Sistu, P. Varley, D. O'Dea, M. Uricár, S. Milz, M. Simon, K. Amende, et al., Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9308–9318.
2. Deng, L., M. Yang, Y. Qian, C. Wang, and B. Wang, CNN based semantic segmentation for urban traffic scenes using fisheye camera. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2017, pp. 231–236.
3. Kumar, V. R., M. Klingner, S. Yogamani, S. Milz, T. Fingscheidt, and P. Mader, Syndistnet: Self-supervised monocular fisheye camera distance estimation synergized with semantic segmentation for autonomous driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 61–71.
4. Shi, H., Y. Li, K. Yang, J. Zhang, K. Peng, A. Roitberg, Y. Ye, H. Ni, K. Wang, and R. Stiefelhagen, FishDreamer: Towards Fisheye Semantic Completion via Unified Image Outpainting and Segmentation. *arXiv preprint arXiv:2303.13842*, 2023.
5. Feng, D., C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 22, No. 3, 2020, pp. 1341–1360.
6. Kirillov, A., K. He, R. Girshick, C. Rother, and P. Dollár, Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9404–9413.
7. Wu, B., A. Wan, X. Yue, and K. Keutzer, Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *2018 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2018, pp. 1887–1893.
8. Valada, A., G. L. Oliveira, T. Brox, and W. Burgard, Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In *2016 International Symposium on Experimental Robotics*, Springer, 2017, pp. 465–477.
9. Sun, Y., W. Zuo, and M. Liu, RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, Vol. 4, No. 3, 2019, pp. 2576–2583.
10. Cordts, M., 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.
11. Geiger, A., P. Lenz, and R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2012, pp. 3354–3361.
12. Wang, S., M. Bai, G. Mattyus, H. Chu, W. Luo, B. Yang, J. Liang, J. Cheverie, S. Fidler, and R. Urtasun, Torontocity: Seeing the world with a million eyes. *arXiv preprint arXiv:1612.00423*, 2016.
13. Huang, X., X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 954–960.

14. Sáez, Á., L. M. Bergasa, E. López-Guillén, E. Romera, M. Tradacete, C. Gómez-Huélamo, and J. Del Egido, Real-time semantic segmentation for fisheye urban driving images based on ERFNet. *Sensors*, Vol. 19, No. 3, 2019, p. 503.
15. Playout, C., O. Ahmad, F. Lecue, and F. Cheriet, Adaptable deformable convolutions for semantic segmentation of fisheye images in autonomous driving systems. *arXiv preprint arXiv:2102.10191*, 2021.
16. Blott, G., M. Takami, and C. Heipke, Semantic segmentation of fisheye images. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.
17. Sistu, G., I. Leang, and S. Yogamani, Real-time joint object detection and semantic segmentation network for automated driving. *arXiv preprint arXiv:1901.03912*, 2019.
18. Ye, Y., K. Yang, K. Xiang, J. Wang, and K. Wang, Universal semantic segmentation for fisheye urban driving images. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2020, pp. 648–655.
19. Chen, X., Y. Yuan, G. Zeng, and J. Wang, Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2613–2622.
20. Tarvainen, A. and H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, Vol. 30, 2017.
21. Yun, S., D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.
22. Schneider, L., M. Jasch, B. Fröhlich, T. Weber, U. Franke, M. Pollefeys, and M. Rätsch, Multimodal neural networks: RGB-D for semantic segmentation and object detection. In *Image Analysis: 20th Scandinavian Conference, SCIA 2017, Tromsø, Norway, June 12–14, 2017, Proceedings, Part I 20*, Springer, 2017, pp. 98–109.
23. Badrinarayanan, V., A. Kendall, and R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 39, No. 12, 2017, pp. 2481–2495.
24. Teichmann, M., M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE intelligent vehicles symposium (IV)*, IEEE, 2018, pp. 1013–1020.
25. Long, J., 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.
26. Erkent, Ö. and C. Laugier, Semantic segmentation with unsupervised domain adaptation under varying weather conditions for autonomous vehicles. *IEEE Robotics and Automation Letters*, Vol. 5, No. 2, 2020, pp. 3580–3587.
27. Kalluri, T., G. Varma, M. Chandraker, and C. Jawahar, Universal semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5259–5270.
28. Huo, X., L. Xie, J. He, Z. Yang, W. Zhou, H. Li, and Q. Tian, ATSO: Asynchronous teacher-student optimization for semi-supervised image segmentation. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 1235–1244.

29. Hung, W.-C., Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018.
30. Novosel, J., P. Viswanath, and B. Arsenali, Boosting semantic segmentation with multi-task self-supervised learning for autonomous driving applications. In *Proc. of NeurIPS-Workshops*, 2019, Vol. 3.
31. Cui, L., X. Jing, Y. Wang, Y. Huan, Y. Xu, and Q. Zhang, Improved Swin Transformer-Based Semantic Segmentation of Postearthquake Dense Buildings in Urban Areas Using Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 16, 2022, pp. 369–385.
32. Zheng, S., J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.
33. Zhang, B., Z. Tian, Q. Tang, X. Chu, X. Wei, C. Shen, et al., Segvit: Semantic segmentation with plain vision transformers. *Advances in Neural Information Processing Systems*, Vol. 35, 2022, pp. 4971–4982.
34. Xie, E., W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, Vol. 34, 2021, pp. 12077–12090.
35. Kumar, V. R., S. Yogamani, H. Rashed, G. Sitsu, C. Witt, I. Leang, S. Milz, and P. Mäder, Omnidet: Surround view cameras based multi-task visual perception network for autonomous driving. *IEEE Robotics and Automation Letters*, Vol. 6, No. 2, 2021, pp. 2830–2837.
36. Zou, Y., Z. Zhang, H. Zhang, C.-L. Li, X. Bian, J.-B. Huang, and T. Pfister, Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020.
37. French, G., S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019.
38. Sohn, K., 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*, Vol. 33, 2020, pp. 596–608.