Team 26: Analysis of Self-Supervised Learning Methods for Urban Scene Segmentation with Adverse Weather Conditions

Anchit Gupta
7035463

Monseej Purkayastha
7047530

Abstract

This project explores the utility of self-supervised learning for urban scene segmentation in adverse weather (low-visibility) conditions with limited labeled data. We assess the effectiveness of pretext tasks, Rotation and Jigsaw, in enhancing U-Net and DeepLabv3 models using Cityscapes and Foggy Cityscapes datasets. Our results show that self-supervised learning with these tasks achieves pixel accuracy comparable to fully supervised methods while reducing the need for labeled data. However, simple pretext tasks have limited success in complex domains like Foggy Cityscapes. These findings underscore the promise of self-supervised learning in data-scarce environments and suggest avenues for future research, including more advanced domain adaptation techniques.

1. Introduction

In this project, we focus on developing a robust urban scene segmentation model that accurately segments objects in images captured under adverse weather conditions in the presence of limited labeled data. Adverse weather, such as fog, rain, and snow, degrades image quality and challenges traditional supervised learning methods that require large, labeled datasets. Collecting and annotating such data is costly and time-consuming.

To address these challenges, we explore Self-Supervised Learning (SSL) to reduce reliance on large labeled datasets and improve model generalization under adverse weather. Unlike Few-Shot Learning (FSL), SSL leverages unlabeled data through pretext tasks, helping the model learn useful representations without explicit supervision. While SSL has shown promise in various domains, its application to urban scene segmentation under adverse weather remains under-explored.

Building on previous research, such as Chen *et al*. 's work on SSL for segmentation [1], our project conducts a comparative analysis of SSL techniques, focusing on their performance in adverse weather. Using the Cityscapes [3] and Foggy Cityscapes [11] datasets, we extend the evaluation of SSL's effectiveness beyond controlled conditions to more challenging scenarios.

2. Related Work

Semantic segmentation has traditionally relied on fully supervised approaches requiring large labeled datasets. Early models like Fully Convolutional Networks (FCNs) [7] and U-Net [10] extended convolutional networks to dense prediction tasks. However, the need for extensive labeled data, especially in complex urban scenes, is a significant limitation. To address this, self-supervised learning methods and synthetic datasets have gained traction. Sakridis *et al*. [11] showed that training with synthetic foggy scenes can enhance the robustness of the model in urban scenarios from the real world in adverse weather. Hoyer *et al*. [4] introduced the DAFormer architecture, using domain-adaptive strategies to align features across different domains, improving generalization. In the 3D space, Liu *et al*. [6] combined active learning with self-training to enhance 3D scene segmentation with minimal supervision. For urban scene understanding, Jiang *et al*. [5] proposed learning relative depth from monocular images without requiring ground-truth depth annotations.

3. Methodology

This section describes the methodology of our project, including dataset selection and splitting, pretext tasks for self-supervised learning, chosen model architectures, loss functions, and the training procedure. We also cover the evaluation metrics used to assess the performance of the model and the domain adaptation strategy to test the robustness of the model under different environmental conditions.

3.1. Datasets

1. The **Cityscapes dataset** [3] is our primary source of urban scene images. It includes more than 5,000 finely annotated images from 50 cities representing diverse

urban environments. The dataset features images captured in favorable weather conditions, serving as a baseline for urban scene segmentation tasks.

2. The **Foggy Cityscapes dataset** [11] is a modified version of the Cityscapes dataset with synthetic fog added to simulate adverse weather conditions. This dataset is essential for evaluating model performance under challenging visibility conditions similar to real-world scenarios in autonomous driving. It includes three variations of images with different fog density levels. For our project, we selected the variation with a fog density parameter of $beta = 0.01$, representing a moderate level of fog that typically occurs, providing a balanced test environment for our models.

3.2. Data splitting

The datasets are divided into training and testing sets. For the self-supervised learning (SSL) phase, 80% of the images (2380) are used as unlabeled data for training the encoder, while the remaining 20% (595) serve as labeled data for fine-tuning. This split assesses SSL methods' effectiveness with limited labeled data. For baseline models, 100% of the images (2975) are used for training. For testing, 500 images are used from each dataset.

3.3. Pretext Tasks for Self-Supervised Learning

1. **Rotation Prediction Task**: Predict the rotation angle of an image among 0°, 90°, 180°, and 270°. This allows the model learn spatial representations and understand urban scene structure, aiding in segmentation.

2. **Jigsaw Puzzle Task**: Reconstruct the original image from shuffled 3x3 patches. This helps to develop an understanding of image region relationships, crucial for accurate segmentation.

3.4. Models

1. **U-Net:** The U-Net model [10] is a CNN architecture for image segmentation, featuring a U-shaped structure with a contracting path for context and an expanding path for precise localization. This design integrates global and local features, making U-Net effective in generating accurate segmentation maps. It is well-suited for semantic segmentation with limited labeled data, and its skip connections preserve fine details, ensuring precise segmentation in challenging conditions. For our project, U-Net's ability to handle varying feature scales and perform well in adverse weather improves its effectiveness in segmenting urban scenes under diverse conditions.

2. **DeepLabv3:** DeepLabv3 [2] is designed for semantic segmentation, employing atrous (dilated) convolutions to capture multi-scale contextual information without increasing computational burden. The model's backbone is a ResNet50 network, augmented with the atrous spatial pyramid pooling (ASPP) module to gather context at multiple scales. Our implementation uses a 5-branch ASPP module with 1x1 convolution, 3x3 convolutions with dilation rates of 6, 12, and 18, and a global pooling block for adaptive average pooling. DeepLabv3 excels in segmenting complex images by capturing fine details and broader contextual information. Its design is advantageous in scenarios with large variations in object scale and appearance. DeepLabv3's strong multi-scale feature extraction capabilities are useful in learning representations of complex objects and scenes at different scales.

3.5. Loss Functions

1. **Pretext Task Loss:** For pretext tasks like rotation prediction and jigsaw puzzle solving, we use **cross-entropy loss** to optimize the models. This loss quantifies the discrepancy between predicted class probabilities and actual class labels. The cross-entropy loss ($\mathcal{L}CE$) for a single sample is defined as:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C} y_c \log(p_c) \tag{1}$$

where (C) is the total number of classes, (y_c) is a binary indicator if the class label (c) is correct, and (p_c) is the predicted probability for class (c). Minimizing this loss helps the model learn robust feature representations.

2. **Segmentation Task Losses:** During fine-tuning, we use a combination of **cross-entropy loss** and **Dice loss** for semantic segmentation:

   (a) **Cross-Entropy Loss:** Each pixel is treated as a separate classification problem, with the cross-entropy loss ($\mathcal{L}_{CE}$) measuring pixel-wise classification accuracy.

   (b) **Dice Loss:** To address the class imbalance problem inherent in segmentation tasks, we incorporate Dice loss [8] ($\mathcal{L}_{Dice}$), which is calculated as

$$\mathcal{L}_{Dice} = 1 - \frac{2|P \cap G|}{|P| + |G|} \tag{2}$$

   where (P) is the predicted segmentation mask and (G) is the ground truth mask. The intersection ($|P \cap G|$) quantifies the overlap between predicted and actual segments, while ($|P|$) and ($|G|$) are the sizes of the predicted and ground truth masks. This loss ensures accurate segmentation

of less frequent objects. By combining cross-entropy and Dice losses, the model balances accurate pixel classification and class imbalance handling. The total loss ($\mathcal{L}total$) is expressed as:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{CE} + \beta\mathcal{L}_{Dice} \quad (3)$$

where, (α) and (β) are hyperparameters that control the trade-off between the two loss components. For our experiments, we have set them to be equal for both losses.

3.6. Training Procedure

Given the project timeline and available computational resources, we decided to train our models for 8 epochs on each pretext task and 100 epochs for supervised fine-tuning. These choices were based on the analysis of the loss curve, which indicated that the loss was approaching saturation at these points. For DeepLabv3, although we intended to train for around 600 epochs (as is done with most implementations), the training process was excessively lengthy and impractical for our needs.

3.7. Evaluation Metrics

To evaluate the performance of our segmentation models, we utilized the following key metrics:

1. **Pixel Accuracy:** It measures the percentage of correctly classified pixels in the image, providing an overall assessment of segmentation performance.

2. **Mean Intersection over Union (mIoU):** Calculates the average IoU across all classes, assessing the overlap between predicted segmentation and ground truth, crucial for complex scenes.

3. **Precision:** It quantifies the ratio of correctly predicted positive pixels to total predicted positive pixels, important for identifying relevant objects and reducing false positives.

4. **Recall:** It measures the ratio of correctly predicted positive pixels to all actual positive pixels in the ground truth, critical to detecting all relevant objects and minimizing false negatives.

These metrics together offer a comprehensive evaluation of the model's segmentation accuracy and its effectiveness in identifying and classifying objects within urban scenes, particularly under challenging conditions.

3.8. Domain Adaptation

The domain adaptation phase evaluates how well a model trained on clear weather conditions (Cityscapes) performs in adverse weather conditions (Foggy Cityscapes). The model is first trained on Foggy Cityscapes with self-supervised pretext tasks, then fine-tuned on Cityscapes data, and finally evaluated on Foggy Cityscapes. A key challenge is that pretext tasks effective in clear conditions may not handle the visual ambiguity of foggy environments well, potentially leading to suboptimal performance. This issue and its impact on cross-domain performance are analyzed in the results section (Tab. 4).

4. Experimental Results and Analysis

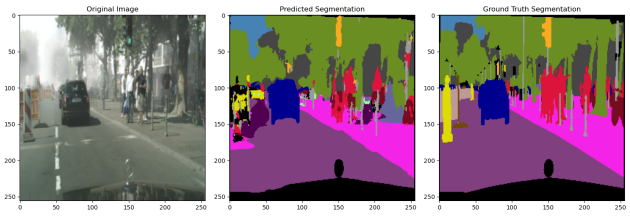
	Model	Accuracy	mIoU	Precision	Recall
Cityscapes	Baseline	88.45%	51.82%	68.91%	62.05%
	Rotate	84.46%	37.13%	52.36%	45.79%
	Jigsaw	85.61%	38.91%	53.35%	48.01%
Foggy	Baseline	88.11%	50.44%	67.65%	59.96%
	Rotate	84.20%	37.12%	49.91%	48.76%
	Jigsaw	85.05%	38.17%	51.29%	48.87%

Table 1. Results for U-Net architecture

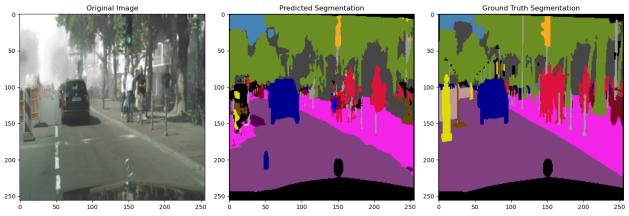
	Model	Accuracy	mIoU	Precision	Recall
Cityscapes	Baseline	80.70%	27.25%	49.30%	34/70%
	Rotate	78.08%	19.43%	36.74%	24.76%
	Jigsaw	78.43%	18.95%	40.65%	24.67%
Foggy	Baseline	79.79%	26.93%	49.57%	34.07%
	Rotate	77.03%	19.49%	40.37%	25.31%
	Jigsaw	77.35%	18.95%	44.28%	24.31%

Table 2. Results for DeepLabv3 architecture

From Tab. 1 and Tab. 2, we observe that SSL shows performance comparable to fully supervised settings. The jigsaw pretext task outperforms the rotation task for both architectures, suggesting that the more complex task helps the model learn scene representations more robustly. This can also be seen by observing the outputs of the rotation and jigsaw pretext tasks in Fig. 1 and Fig. 2. Surprisingly, DeepLabv3 performs worse than U-Net, likely due to fewer training epochs (100 instead of the typical 600 for high-quality results).

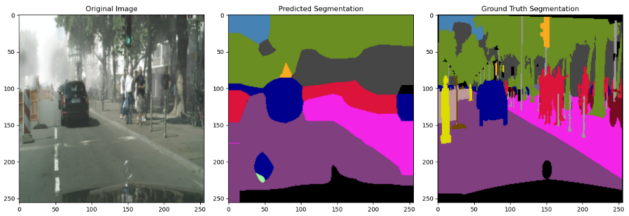


(a) Rotation Pretext

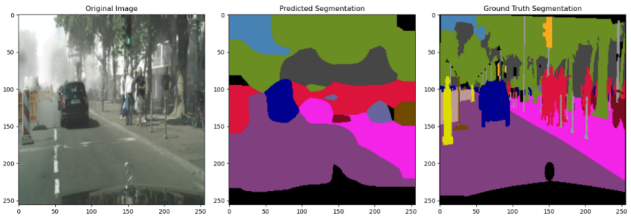


(b) Jigsaw Pretext

Figure 1. Sample outputs for U-Net for Foggy Cityscapes



(a) Rotation Pretext



(b) Jigsaw Pretext

Figure 2. Sample outputs for DeepLabv3 for Foggy Cityscapes

Model	Training Time (in hrs.)
Baseline	≈ 13
Rotate + fine-tuning	≈ 6
Jigsaw + fine-tuning	≈ 7

Table 3. Training time under different settings

We also noticed a significant difference in training time (Tab. 3) for baseline models vs. models trained with self-supervised learning & fine-tuned using supervised learning.

	Model	Accuracy	mIoU	Precision	Recall
U-Net	Baseline	88.11%	50.44%	67.65%	59.96%
	Rotate	70.36%	23.03%	42.65%	33.49%
	Jigsaw	76.56%	26.96%	46.77%	36.67%
DeepLabv3	Baseline	79.79%	26.93%	49.57%	34.07%
	Rotate	62.05%	17.47%	38.10%	26.45%
	Jigsaw	72.43%	24.57%	50.42%	33.42%

Table 4. Results for U-Net and DeepLabv3 architectures for Domain Adaptation

The domain adaptation results in Tab. 4 reveals challenges in generalizing from clear to adverse weather. U-Net and DeepLabv3 models, trained on Foggy Cityscapes and fine-tuned on Cityscapes, showed a notable performance drop when tested in Foggy Cityscapes, especially in mIoU and Pixel Accuracy. This highlights difficulties in transferring knowledge between different domains. However, using the jigsaw pretext task for domain adaptation consistently outperformed the rotation task across all metrics.

5. Conclusions

This project aimed to develop and evaluate a robust urban scene segmentation model for diverse and challenging weather conditions. Using U-Net and DeepLabv3 architectures, models were trained with self-supervised learning techniques and fine-tuned with labeled data. The results showed good performance in clear weather, but significant challenges in domain adaptation to adverse weather such as fog, leading to reduced segmentation accuracy.

Pretext tasks like rotation prediction and jigsaw puzzle solving helped the model learn useful features but did not fully capture the complexities of foggy conditions. However, the model showed reasonable generalizability in identifying key objects under adverse conditions, with the jigsaw task outperforming rotation. Incorporating Pretext-Invariant Representation Learning (PIRL) [9] and data augmentation with GANs and diffusion models could further enhance generalization.

The project underscores the potential and limitations of current urban scene segmentation in adverse weather. Future work should focus on improving domain adaptation and exploring advanced self-supervised tasks to boost model robustness and generalizability in real-world scenarios.

References

[1] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Ekin D Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 695–714. Springer, 2020. 1

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. 2

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[4] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Domain adaptive and generalizable network architectures and training strategies for semantic image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 46(1):220–235, 2021. 1

[5] Pingbo Jiang, Xiangyang Ji, Qidi Wu, and Siyuan Huang. Self-supervised relative depth learning for urban scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12623–12632, 2021. 1

[6] Shaoshuai Liu, Chi Zhang, Chi Xu, Yiyi Zhao, Chen Change Loy, and Dahua Lin. Active self-training for weakly supervised 3d scene semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1567–1576, 2021. 1

[7] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 1

[8] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016. 2

[9] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. 1, 2

[11] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 1, 2