CSE Semester Project

SAM vs. U-net: A Comparative Analysis for Advanced Road Segmentation

Written by:
Maurer Thamin

Supervised by:
Professor Nikolaos Geroliminis
PhD Candidate Yura Tak

January 10, 2024

Contents

1 Introduction

In today's crowded urban landscapes, Unmanned Aerial Vehicles (UAVs) have emerged as powerful tools for aerial surveillance, providing real-time insights into traffic patterns, monitoring public safety, and ensuring overall urban efficiency. Despite their potential, the accurate identification and tracking of vehicles in UAV-based surveillance footage encounter challenges, particularly in the context of dynamic road scenes.

This project focuses on the critical task of road segmentation, aiming to enhance the effectiveness of vehicle detection in urban environments. The primary motivation behind this shift lies in the recognition that precise road segmentation serves as a foundational element for robust vehicle detection, enabling better navigation through complex urban landscapes.

To achieve this, we employed two cutting-edge deep learning models, SAM (Segment Anything Model) and U-net architecture. SAM, developed by Meta AI, is a versatile image segmentation model that incorporates spatial attention modules to focus on salient features and refine segmentation boundaries. This makes it well-suited for handling complex road scenes. The U-net architecture, a widely used convolutional neural network (CNN) for image segmentation, is known for its ability to capture fine details and produce high-quality segmentation masks.

2 Related Work

2.1 Early Approaches to Road Segmentation

In the early stages of road segmentation, a variety of techniques were employed, many of which leveraged the power of deep learning. Mahmud et al. (2021) [1] utilized UAV images in conjunction with the DeepLab V3+ Semantic Segmentation Model for road image segmentation. This approach demonstrated the potential of UAV images in enhancing the accuracy of road segmentation.

Around the same time, Ojha et al. (2021) [2] proposed an instance segmentation approach for vehicle detection using Mask R-CNN. Although their focus was on vehicle detection, the underlying principle of instance segmentation laid a foundation that could be adapted for road segmentation tasks.

In a similar vein, Sun et al. (2021) [3] developed a road segmentation algorithm based on a simplified version of Mask R-CNN, specifically for GF-3 SAR images. This work highlighted the adaptability of Mask R-CNN for different types of images and its effectiveness in road segmentation.

Kherraki et al. (2021) [4] explored traffic scene semantic segmentation using several deep convolutional neural networks. Their work underscored the potential of using a combination of networks to enhance the accuracy of semantic segmentation in traffic scenes.

Lastly, the use of U-Net, a type of convolutional network, also gained popularity in road segmentation tasks. Hao et al. (2023) [5] presented a multi-objective semantic segmentation algorithm based on an improved U-Net network. Their work emphasized the role of U-Net in improving the recognition efficiency of various types of features in the construction zone of transportation facilities.

These early approaches provided valuable insights and set the stage for more advanced techniques in road segmentation.

2.2 U-Net

The U-Net model is a popular architecture used for image segmentation tasks. It was originally proposed for medical image segmentation [6]. The architecture of the U-Net model is characterized by a contracting path to capture context and a symmetric expanding path that enables precise localization.

2.2.1 U-Net Architecture Overview

The U-Net architecture follows an encoder-decoder cascade structure. The encoder gradually compresses information into a lower-dimensional representation, and then the decoder decodes this information back to the original image dimension. This gives the architecture an overall U-shape, which leads to the name U-Net.

One of the prominent features of the U-Net architecture is the skip connections, which enable the flow of information from the encoder side to the decoder side, enabling the model to make better predictions.

2.2.2 Image Segmentation with U-Net

In Image Segmentation, the model is asked to classify each pixel in the image to the object category it represents. This can be viewed as pixel-level image classification and is a much harder task than simple image classification, detection, or localization. The model must automatically determine all objects and their precise location and boundaries at a pixel level in the image. Thus image segmentation provides an intricate understanding of the image and is widely used in medical imaging, autonomous driving, robotic manipulation, etc.

2.3 SAM

The Segment Anything Model (SAM) is a model developed by Meta AI Research for image segmentation tasks [7].

2.3.1 SAM Architecture Overview

SAM uses an image encoder that produces a one-time embedding for the image, while a lightweight encoder converts any prompt into an embedding vector in real-time. These two information sources are then combined in a lightweight decoder that predicts segmentation masks.

2.3.2 Image Segmentation with SAM

SAM is capable of segmenting any object on a certain image. It produces high-quality object masks from input prompts such as points or boxes, and it can be used to generate masks for all objects in an image. It has been trained on a dataset of 11 million images and 1.1 billion masks, and has strong zero-shot performance on a variety of segmentation tasks.
SAM has learned a general notion of what objects are, and it can generate masks for any object in any image or any video, even including objects and image types that it had not encountered during training. SAM is general enough to cover a broad set of use cases and can be used out of the box on new image "domains" without requiring additional training.

2.4 Recent Advances

Recent advances in road segmentation have been driven by the development of deep learning models. For instance, the Seg-Road model, a segmentation network for road extraction based on Transformer and CNN with Connectivity Structures, has been proposed [8]. This model uses a transformer structure to extract long-range dependency and global contextual information to improve the fragmentation of road

segmentation. It also uses a convolutional neural network (CNN) structure to extract local contextual information to improve the segmentation of road details.
Another significant advancement is the development of efficient deep models for monocular road segmentation [9]. These models exploit recent advances in semantic segmentation via CNNs and provide a good trade-off between segmentation quality and runtime.
Another recent advancement in road segmentation is the use of Generative Adversarial Networks (GANs). GANs have been used to generate synthetic road images, which are then used to augment the training data for road segmentation models. This approach has been shown to improve the robustness of these models, particularly in scenarios where the available training data is limited.
In addition, there has been a growing interest in the use of multi-modal data for road segmentation. This involves the use of data from different sensors (e.g., LiDAR, radar, and cameras) to improve the accuracy of road segmentation. Multi-modal data provides a more comprehensive understanding of the road environment, which can be particularly useful in complex scenarios such as urban environments or adverse weather conditions.

2.5 Critical Evaluation

The U-Net model has shown its potential as a robust and accurate tool for road segmentation [10] [11]. It has been evaluated on the Massachusetts Roads Dataset [12] with the estimation of numerous parameters such as filter stride, learning rate, training epochs, data size, and various augmentation techniques. The results demonstrate that the U-Net model is computationally efficient and achieves comparable segmentation results.
On the other hand, the SAM model, while showing promise in its ability to segment any object in an image, suggests the need for further improvements to enhance segmentation performance [13]. The pretrained SAM architecture has been found to be less robust in certain segmentation tasks compared to the U-Net model.
While the U-Net and SAM models have shown promising results in road segmentation, they also have their limitations. For instance, the U-Net model can struggle with segmenting small or thin structures in images due to the loss of resolution in the encoding process. On the other hand, the SAM model, while capable of segmenting any object, can sometimes produce less precise segmentation masks compared to other models.
Furthermore, both models require a significant amount of computational resources, which can be a limiting factor in real-time applications. There is also a need for large, annotated datasets for training these models, which can be time-consuming and expensive to produce.
In conclusion, while there have been significant advancements in road segmentation, there is still room for improvement. Future research could focus on addressing these limitations and developing more efficient and accurate models for road segmentation.

3 Datasets

3.1 Introduction

Advancements in road segmentation demand specialized datasets that capture real-world challenges, such as occluded vehicles. In this comparative analysis of SAM and U-net for advanced road segmentation, we utilized three key datasets: UAVDT [14], UAVid [15] and a dataset from EPFL [16].

3.2 UAVID

The UAVid dataset is a high-resolution UAV semantic segmentation dataset introduced by Lyu et al [15]. Here are some key details about the dataset:

- **Content :** The UAVid dataset consists of 30 video sequences, capturing 4K high-resolution images in slanted views. These images provide both top and side views of objects, offering more information for object recognition. In total, 300 images have been densely labeled with 8 classes for the semantic labeling task.
- The 8 classes are as follows :
  1. Building
  2. Road
  3. Tree
  4. Low vegetation
  5. Moving car
  6. Static car
  7. Human
  8. Background clutter
- **Challenges :** The dataset introduces new challenges to the field of semantic segmentation, including large scale variation, moving object recognition, and temporal consistency preservation. These challenges make the UAVid dataset a valuable resource for developing and testing new algorithms and models in the field of semantic segmentation.
- **Usage :** The UAVid dataset has been used for tasks such as semantic segmentation and scene segmentation. For instance, models like LSKNet-S and UNetFormer have been used for semantic segmentation and scene segmentation tasks on this dataset.
- **Comparison with Other Datasets :** The UAVid dataset serves as a complement to other semantic segmentation datasets. While datasets like Cityscapes [17] and CamVid [18] capture side views of objects with a camera mounted on a driving car, and some datasets capture top views of objects from airborne or satellite images, the UAVid dataset captures urban scenes from an oblique Unmanned Aerial Vehicle (UAV) perspective.

Figure 1: Representative image from the UAVid dataset.

3.3 EPFL Dataset

This dataset is designed for the task of classifying satellite images, specifically separating 16x16 blocks of pixels between roads and the rest. The dataset contains the design and implementation of two convolutional neural networks. The training set consists of 100 satellite images (400x400) with their respective ground truth, and the testing set consists of 50 satellite images (608x608).

Figure 2: Representative image from the EPFL dataset.

3.4 UAVDT

The UAVDT dataset is a large-scale benchmark for object detection and tracking, specifically designed for Unmanned Aerial Vehicles (UAVs). Here are some additional details about the dataset :

- **Dataset Size :** The dataset consists of about 80,000 representative frames from 10 hours of raw videos.
- **Object of Interest :** The objects of interest in this benchmark are vehicles.
- **Annotations :** The frames are manually annotated with bounding boxes and some useful attributes, such as vehicle category and occlusion.
- **Video Sequences :** The UAVDT benchmark consists of 100 video sequences, which are selected from over 10 hours of videos taken with a UAV platform at a number of locations in urban areas.
- **Scenes :** The videos represent various common scenes including squares, arterial streets, toll stations, highways, crossings, and T-junctions.
- **Resolution and Frame Rate :** The videos are recorded at 30 frames per second (fps), with the JPEG image resolution of 1080 $\times$ 540 pixels.

The UAVDT dataset has been used in various research projects, including training the YOLOv5 network [19], and it has been compared against other advanced trackers. The current state-of-the-art model on the UAVDT dataset for object detection is PRB-FPN [20].
The UAVDT dataset presents new challenges for object detection and tracking, such as high density, small object, and camera motion. These challenges make the dataset a valuable resource for developing and testing new algorithms in the field of object detection and tracking.

3.5 Detailled Analysis of UAVDT

3.5.1 Dataset Processing

The UAVDT is renowned for its rich urban scenarios captured by UAVs. To create a subset highlighting vehicles with occlusions, a systematic approach was taken. Bounding boxes delineating vehicles with varying degrees of occlusion—classified as small, medium, and large—were meticulously extracted from UAVDT images.

3.5.2 Dataset Overview

The resulting dataset offers a diverse set of images, each featuring precisely one vehicle with an occlusion. This intentional curation provides nuanced insights into the challenges associated with occlusions in real-world road segmentation tasks.

Characteristics of the Newly Created Dataset

- **Image Variety:** The dataset comprises images capturing vehicles of varying sizes, reflecting the diversity of occlusion scenarios in urban environments.
- **Occlusion Classes:** Vehicles are categorized based on the degree of occlusion—small, medium, or large—offering a nuanced understanding of occlusion challenges in different contexts.

Statistics

Category	Count
Occluded Frames	83,325
Non-Occluded Frames	175,778
Number of Vehicles	745
Non-Occluded Vehicles	684
Small Occlusion	64,004
Medium Occlusion	10,025
Large Occlusion	9,296

Histogram of Occlusion Levels

A crucial aspect of the dataset is the distribution of occlusion levels per vehicle. Figure 3 provides a histogram illustrating the number of occluded frames associated with each vehicle in the dataset. This granular representation allows for a deeper understanding of how different vehicles are affected by occlusion.

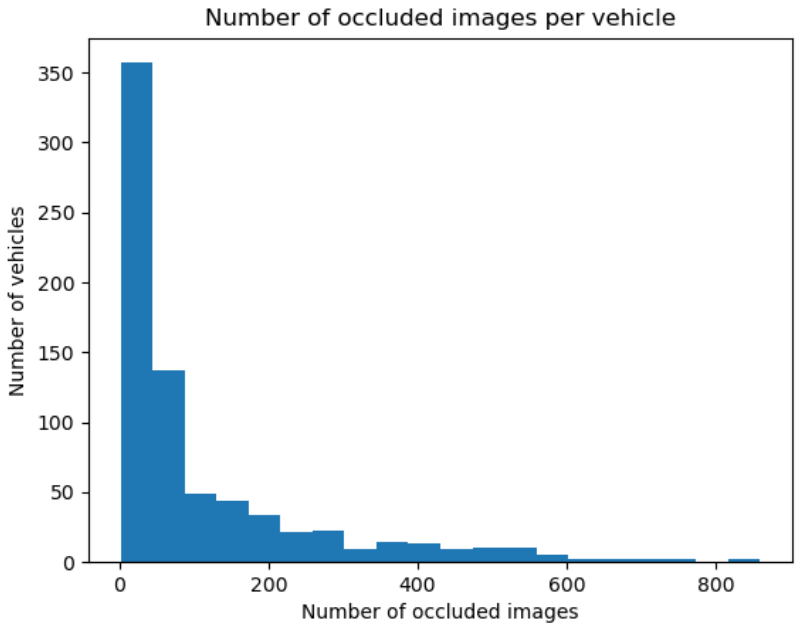


Figure 3: Histogram illustrating the number of occluded frames per vehicle in the dataset.

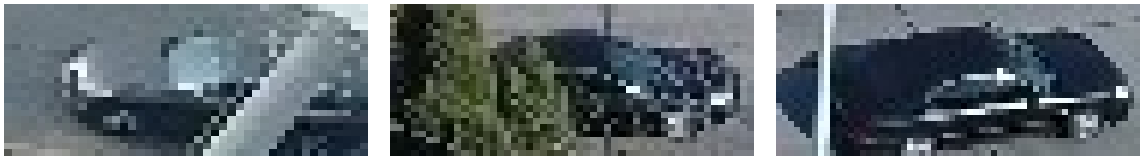
The histogram on figure 3is right-skewed, emphasizing that the majority of vehicles experience a relatively low number of occluded frames, with over 50% having fewer than 200 occluded frames.
While there is a small proportion of vehicles with a higher count of occluded frames, notably beyond 600 occluded frames, it represents only a very small proportion of vehicles.
Key Points :

- Right-skewed distribution: Majority experience low occlusion.
- Median of 150 occluded frames: Half of the vehicles have fewer than 150 occluded frames.
- Limited instances with over 600 occluded frames.

Visualization

To provide a glimpse into the dataset's composition, Figure 4 displays a representative image from the UAVDT dataset, showcasing its intricate urban scenes. Additionally, smaller images of occluded vehicles, highlighting various occlusion scenarios, are presented in Figure 5.

Figure 4: Representative image from the UAVDT dataset.



(a) Medium occlusion level. (b) Big occlusion level. (c) Small occlusion level

Figure 5: Sample images from the newly created dataset, showcasing different occlusion scenarios.

4 Methodology

4.1 U-Net

4.1.1 Data Preprocessing

For data preprocessing, the combination of UAVID and EPFL datasets provided a diverse range of road scenarios. The introduction of rotations at multiple angles, random rotations, and flipped images aimed to enhance the model's ability to handle variations in road orientations. The addition of Gaussian and salt & pepper noise simulated real-world noise in the data.
The use of 96x96 patches allowed the model to learn from smaller regions, and the inclusion of black borders ensured the correct handling of edge cases during training. Ground truth labels were treated as binary images, and the class reassignment for the UAVID dataset facilitated a uniform binary classification. Rounding the ground truth values improved model stability by mitigating the impact of borderline pixel values.

4.1.2 Model Training Strategies

During model training, the choice of binary crossentropy as the loss function aligns well with the binary nature of road segmentation. The Adam optimizer with a learning rate of 1e-4 provided a suitable optimization algorithm. The inclusion of the F1 metric allowed a more comprehensive evaluation beyond accuracy.
The 80% training and 20% validation data split ensured a robust training process while allowing for effective model evaluation on unseen data.

4.1.3 Post-processing Techniques

Post-processing involved fine-tuning predictions to improve segmentation accuracy. The initial predictions on patches provided a localized understanding of road segments. Experimentation with different thresholds enabled the classification of predicted values into road or non-road areas. The final step of image reconstruction from patches produced coherent and visually accurate segmentation maps.

4.1.4 Computational Resources

The U-Net model was trained on an Intel Core i7 11th gen CPU. The training time of approximately 20 hours reflects the computational resources required to optimize the model for road segmentation. The choice of a CPU for training demonstrates the feasibility of achieving meaningful results without the need for specialized hardware.

4.1.5 Validation and Evaluation Metrics

Validation of the U-Net model involved assessing its accuracy and F1 score on the validation set. The F1 score, considering both precision and recall, provided insights into the model's ability to correctly identify road pixels while minimizing false positives and false negatives. Although hyperparameter tuning was not extensively

performed, experimenting with different batch sizes allowed for an exploration of trade-offs between computational efficiency and model performance.

4.2 SAM

4.2.1 Data Preprocessing

For data preprocessing with SAM, computational resource limitations influenced the augmentation strategy. Due to constraints, SAM was trained without extensive data augmentation. The images were converted to grayscale, and to facilitate model training, they were split into 256x256 patches.

4.2.2 Model Training Strategies

SAM leverages a combination of loss functions and optimization algorithms to achieve impressive performance in image segmentation.

Loss Functions :

- **Cross-Entropy Loss:** Measures the difference between predicted masks and ground truth masks, penalizing the model for assigning incorrect labels.
- **Smooth Dice Loss:** Specifically designed for segmentation tasks, less sensitive to class imbalance, combining Dice similarity coefficient (DSC) and cross-entropy loss.
- **Hausdorff Distance Loss:** Measures distance between boundaries of predicted masks and ground truth masks, crucial for tasks requiring precise boundary delineation.

Optimization Algorithms :

- **AdamW Optimizer:** Variant of Adam optimizer with weight decay to prevent overfitting, widely used for deep learning models.
- **Gradient Clipping:** Technique used to prevent gradient explosion during optimization, ensuring stable learning.

SAM was fine-tuned using a pretrained model, adapting it to the specific road segmentation task.

4.2.3 Post-processing Techniques

Post-processing for SAM involved threshold adjustments to optimize predictions on patches. Different thresholds were tested to classify predictions into road or non-road. The final step involved reconstructing entire images from the processed patches.

4.2.4 Computational Resources

SAM's model required substantial computational resources. Training was conducted on the LUTS server, and the process took approximately 10 hours for 10 epochs. The need for a powerful server underscores the resource-intensive nature of training sophisticated segmentation models.

4.2.5 Validation and Evaluation Metrics

SAM is evaluated using a diverse set of metrics to comprehensively assess its performance in image segmentation.
Pixel-Level Metrics :

- **Mean Intersection over Union (mIoU):** Measures the average of IoU for each class, indicating the overall similarity between predicted and ground truth masks.
- **Dice Similarity Coefficient (DSC):** Assesses similarity between predicted and ground truth masks based on the proportion of correctly classified pixels.

Boundary-Aware Metrics :

- **Hausdorff Distance (HD):** Measures the maximum distance between predicted and ground truth boundaries, emphasizing boundary alignment.
- **Variation of Information (VI):** Quantifies uncertainty in segmentation by measuring mutual information between predicted and ground truth masks.

Additionally, SAM is evaluated with task-specific metrics relevant to its application, ensuring a nuanced understanding of its performance in specific domains.

5 Results and Discussion

5.1 Performance of U-Net

The training process of the U-Net model is visualized in Figure 6, showcasing the evolution of loss and F1-Score on both the training and validation sets.

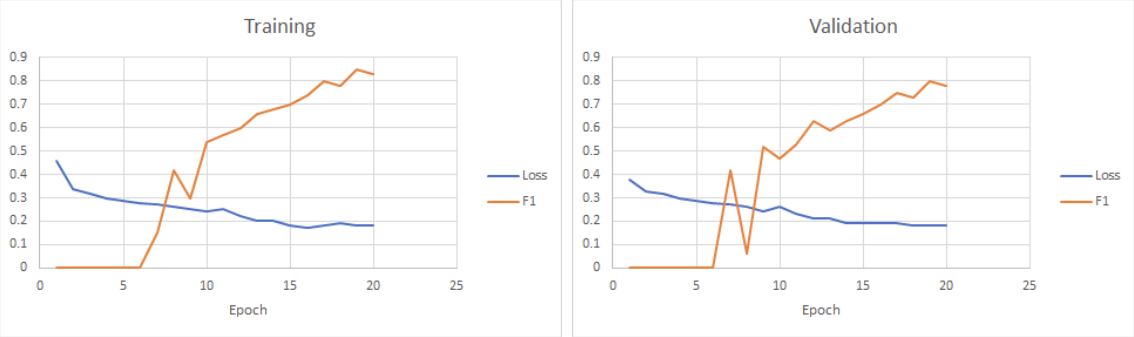


Figure 6: Loss and F1-Score during training and validation.

The U-Net model demonstrates a commendable F1-score of 87%, indicating a well-balanced trade-off between recall and precision. Figure 7 illustrates the model's accurate prediction on a test image from the EPFL dataset.

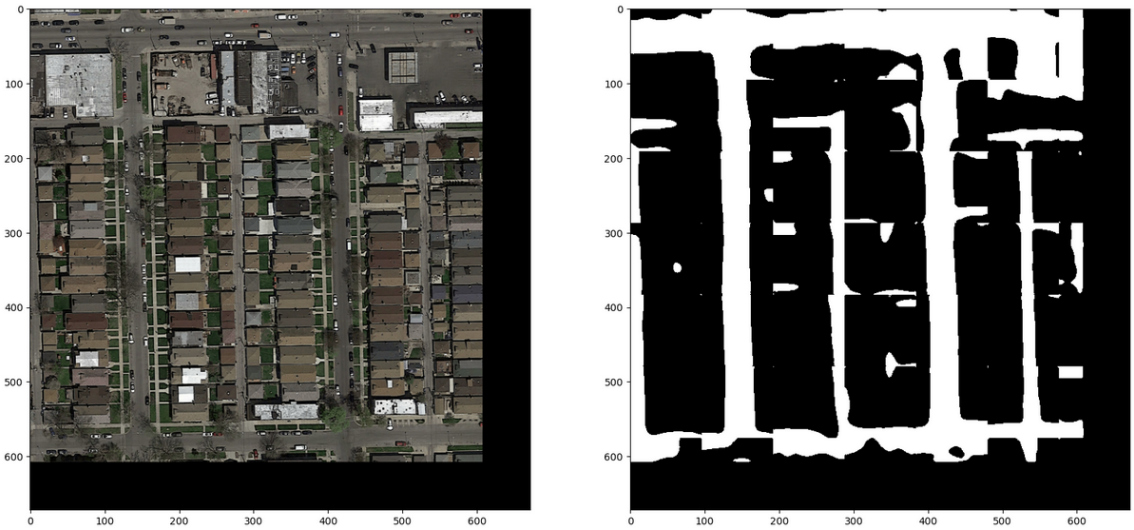


Figure 7: Model output on a test image of the EPFL dataset.

Figures 8, 9, and 10 display the model's predictions on images not used during training.

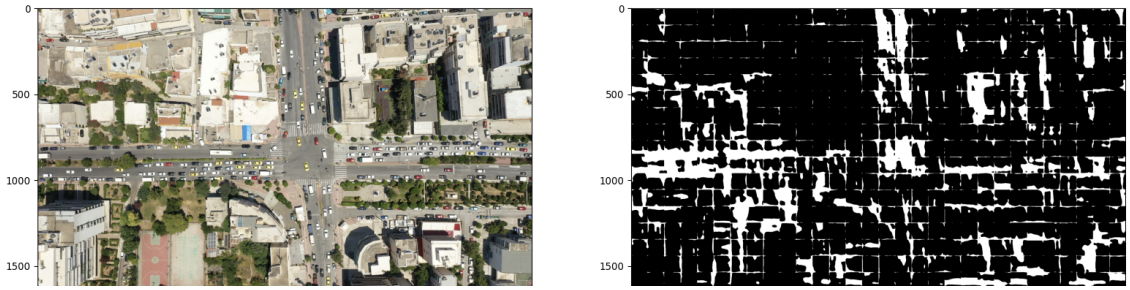


Figure 8: Model output on the Galatsi image.

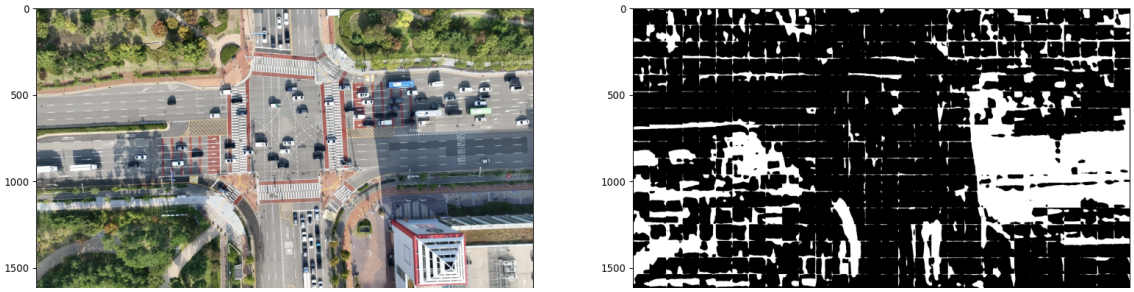


Figure 9: Model output on the SongDo image.

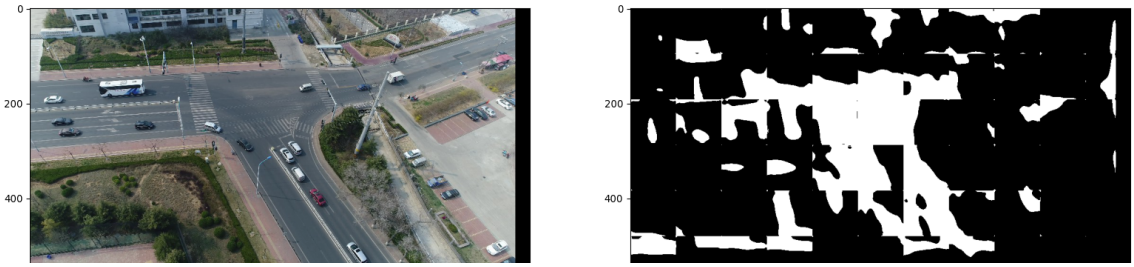


Figure 10: Model output on a test image of the UAVDT dataset.

While the U-Net model excels in certain areas, it faces challenges in generalization. In Figure 8, the model accurately detects roads in specific regions but struggles with others. Notably, in Figures 9 and 10, the model's performance is affected by the increased width of roads not encountered during training. Additionally, Figure 9 reveals a limitation where shadows from various sources contribute to misclassifications.
These observations emphasize the need for enhanced generalization and robustness in handling diverse road characteristics for improved model performance.

5.2 Performance of SAM

Figure 11 presents the SAM model's prediction on a test image from the UAVid dataset, revealing notable noise and improper segmentation of the road.

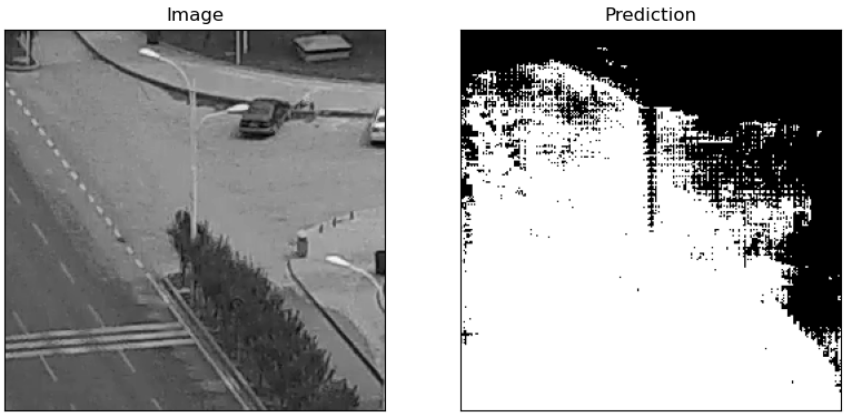


Figure 11: Model output on a test image of the UAVid dataset.

Figures 12, 13, and 14 showcase SAM's predictions on images not included during training.

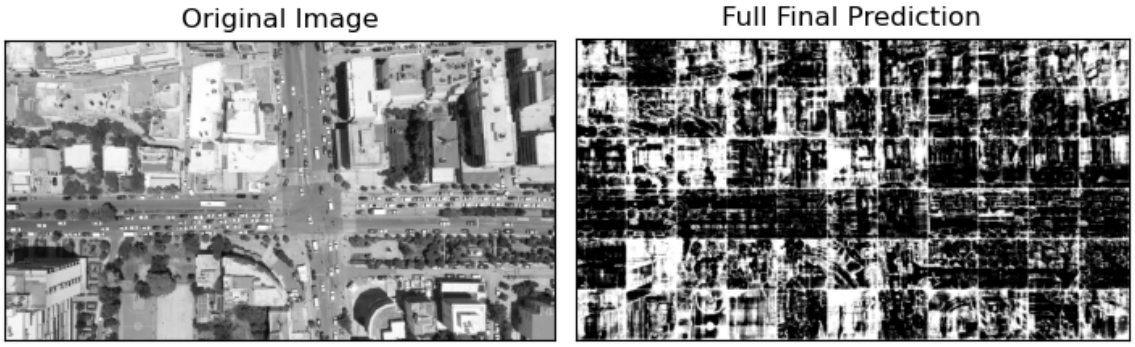


Figure 12: Model output on the Galatsi image..

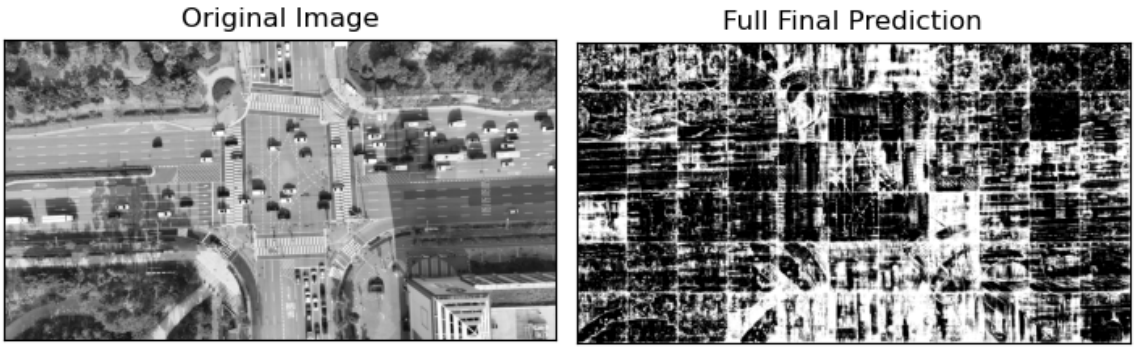


Figure 13: Model output on the SongDo image.

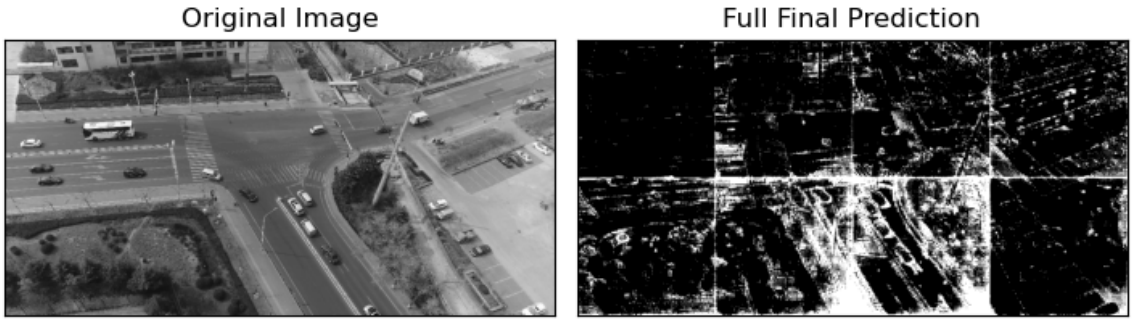
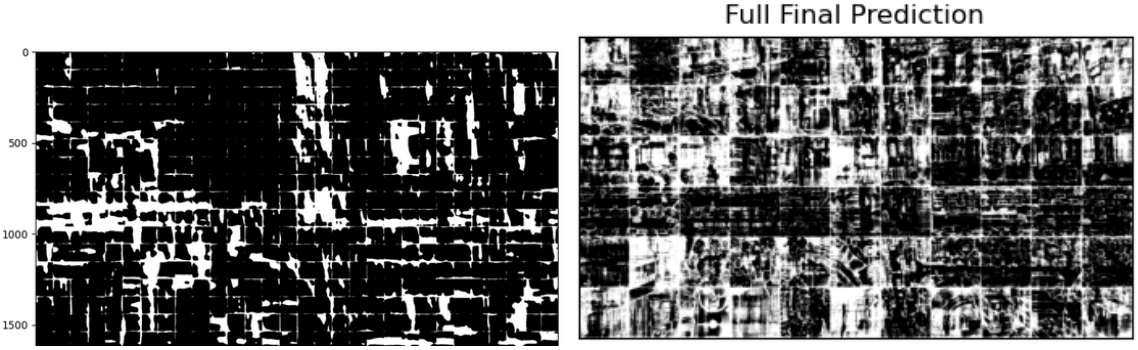


Figure 14: Model output on a test image of the UAVDT dataset.

While SAM demonstrates proficiency in certain aspects, it faces challenges in adapting to road segmentation. Notably, due to computational resource limitations, extensive data augmentation could not be applied to the training dataset. Data augmentation techniques, such as rotations, flips, and variations in lighting conditions, play a crucial role in enhancing a model's ability to generalize to diverse scenarios. The absence of extensive data augmentation may have contributed to the model's difficulty in handling certain road characteristics. In Figure 12, where the model detects a change in the middle right but misclassifies it, the limited diversity in the training data might have played a role. Similarly, the misclassification of the road in the middle bottom of Figure 14 could be attributed, in part, to the constrained variability in the training dataset.
These limitations highlight the importance of access to ample computational resources for comprehensive data augmentation, allowing models like SAM to better generalize across a broader range of road scenarios.

5.3 Comparative Analysis

In this section, we conduct a comparative analysis of the U-Net and SAM models in terms of their performance on road segmentation. Figure 15 presents a visual comparison of their outputs on a sample test image.



(a) U-Net (b) SAM

Figure 15: Comparison of U-Net and SAM on road segmentation..

Both models exhibit strengths and limitations in different scenarios. U-Net demonstrates commendable performance in accurately delineating roads, as shown in Fig-

ure 7. However, its generalization capabilities face challenges when encountering wider roads, as observed in Figures 9 and 10. On the other hand, SAM struggles with noise and improper segmentation, as evident in Figure 11. Despite this, SAM excels in certain areas, such as detecting changes in the middle right of the Galatsi image (Figure 12).

5.4 Challenges and Future Work

For both U-Net and SAM, several challenges and areas for future improvement are identified.

1. **Limited Generalization:** Both models show challenges in generalizing to diverse road characteristics. U-Net struggles with wider roads not encountered during training, while SAM faces difficulties in handling noise and subtle road features.

2. **Enhanced Data Augmentation:** Given the computational resource limitations for SAM, exploring more efficient data augmentation techniques becomes crucial. Augmenting the training dataset with a broader range of scenarios, lighting conditions, and road characteristics can enhance model generalization.

3. **Transfer Learning:** Leveraging pre-trained models on large datasets for transfer learning could enhance the initial training of both U-Net and SAM. Fine-tuning on the specific road segmentation task can capitalize on the knowledge gained from broader image datasets.

4. **Ensemble Approaches:** Combining predictions from multiple models through ensemble methods may mitigate the weaknesses of individual models. Ensemble strategies could enhance overall robustness and accuracy in road segmentation.

Addressing these challenges and exploring these avenues will contribute to the development of more robust and accurate models for road segmentation in aerial images.

6 Conclusion

In the pursuit of accurate road segmentation from aerial images, this exploration dived into the utilization of U-Net and the recently introduced SAM model. Both approaches demonstrated strengths and faced challenges, providing valuable insights into the complexities of this task.

6.1 Performance Overview

U-Net showcased commendable results, achieving an 87% F1-score. Its ability to accurately delineate roads was evident in various scenarios, particularly in well-encountered road configurations. However, challenges emerged in generalizing to diverse road characteristics, particularly wider roads and regions with complex lighting conditions or shadows.
SAM, while presenting proficiency in certain aspects, struggled with noise and improper segmentation. Computational resource limitations constrained the extent of data augmentation, impacting SAM's adaptability to a broader range of road scenarios. Notably, SAM excelled in detecting changes and subtle features but struggled in scenarios demanding robust noise handling.

6.2 Challenges and Future Directions

The limitations identified point to promising avenues for improvement. Both models struggled with limited generalization, emphasizing the importance of enriching training datasets with diverse road characteristics. Enhanced data augmentation techniques, transfer learning from pre-trained models, and the exploration of ensemble approaches stand out as potential strategies to boost performance.

6.3 Final Reflection

Experience with U-Net and SAM has provided a foundation for future advancements in road segmentation. A deeper understanding of their performance has guided improvements in strategies, the use of more diverse datasets, and the exploration of new techniques. As computers and methods become more advanced, the development of stronger and more flexible models for aerial image segmentation is within reach.
This exploration contributes to the overall knowledge in computer vision, paving the way for improvements that enhance road segmentation accuracy and adaptability to real-world complexities.

References

[1] Mat Nizam Mahmud et al. "Road Image Segmentation using Unmanned Aerial Vehicle Images and DeepLab V3+ Semantic Segmentation Model". In: *2021 11th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*. 2021, pp. 176–181. DOI: 10.1109/ICCSCE52189.2021.9530950.

[2] Apoorva Ojha, Satya Prakash Sahu, and Deepak Kumar Dewangan. "Vehicle Detection through Instance Segmentation using Mask R-CNN for Intelligent Vehicle System". In: *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. 2021, pp. 954–959. DOI: 10.1109/ICICCS51141.2021.9432374.

[3] Zengguo Sun, Hui Geng, and Bai Jia. "Road Segmentation Algorithm Based on Simplified Mask R-CNN for GF-3 SAR Images". In: *2021 SAR in Big Data Era (BIGSARDATA)*. 2021, pp. 1–4. DOI: 10.1109/BIGSARDATA53212.2021.9574293.

[4] Amine Kherraki, Muaz Maqbool, and Rajae El Ouazzani. "Traffic Scene Semantic Segmentation by Using Several Deep Convolutional Neural Networks". In: *2021 3rd IEEE Middle East and North Africa COMMunications Conference (MENACOMM)*. 2021, pp. 1–6. DOI: 10.1109/MENACOMM50742.2021.9678270.

[5] Xuejie Hao et al. "A Multi-Objective Semantic Segmentation Algorithm Based on Improved U-Net Networks". In: *Remote Sensing* 15.7 (2023). ISSN: 2072-4292. DOI: 10.3390/rs15071838. URL: https://www.mdpi.com/2072-4292/15/7/1838.

[6] Thomas Brox Olaf Ronneberger Philipp Fischer. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *arXiv preprint arXiv:1505.04597* (2015).

[7] Qi Fan et al. *Stable Segment Anything Model*. 2023. arXiv: 2311.15776 [cs.CV].

[8] Nitish Katal Kushagra Pal Piyush Yadav. "RoadSegNet: a deep learning framework for autonomous urban road detection". In: *Journal of Engineering and Applied Science* (2022). URL: https://jeas.springeropen.com/articles/10.1186/s44147-022-00162-9.

[9] Fernando Santos Osório Felipe Manfio Barbosa. "A Threefold Review on Deep Semantic Segmentation: Efficiency-oriented, Temporal and Depth-aware design". In: *arXiv preprint arXiv:2303.04315* (2023). URL: https://arxiv.org/abs/2303.04315.

[10] Sangpil Kim Eugenio Culurciello Adam Paszke Abhishek Chaurasia. "ENet: A deep neural network architecture for real-time semantic segmentation". In: *arXiv preprint arXiv:1606.02147* (2016). URL: https://arxiv.org/abs/1606.02147.

[11] Manoj Kumar Sachan Vidhi Chaudhary Preetpal Kaur Buttar. "Satellite imagery analysis for road segmentation using U-Net architecture". In: *The Journal of Supercomputing* (2022). URL: https://link.springer.com/article/10.1007/s11227-022-04379-6.

[12] Volodymyr Mnih. "Machine Learning for Aerial Image Labeling". PhD thesis. University of Toronto, 2013.

[13] "Road Segmentation using U-Net architecture". In: *2020 IEEE International conference of Moroccan Geomatics (Morgeo)*. 2020. URL: https://ieeexplore.ieee.org/abstract/document/9121887.

[14] Dawei Du et al. "The unmanned aerial vehicle benchmark: Object detection and tracking". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 370–385. URL: https://sites.google.com/view/grli-uavdt/%E9%A6%96%E9%A1%B5.

[15] Ye Lyu et al. "UAVid: A semantic segmentation dataset for UAV imagery". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 165 (2020), pp. 108–119. URL: https://uavid.nl/.

[16] EPFL ML. *EPFL ML Road Segmentation*. https://www.aicrowd.com/challenges/epfl-ml-road-segmentation. Accessed: 2024-01-03. 2023.

[17] Marius Cordts et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[18] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. "Semantic object classes in video: A high-definition ground truth database". In: *Pattern Recognition Letters* (2008).

[19] Glenn Jocher. *Ultralytics YOLOv5*. Version 7.0. 2020. DOI: 10.5281/zenodo.3908559. URL: https://github.com/ultralytics/yolov5.

[20] Jun-Wei Hsieh Yong-Sheng Chen Ping-Yang Chen Ming-Ching Chang. "Parallel Residual Bi-Fusion Feature Pyramid Network for Accurate Single-Shot Object Detection". In: *IEEE Transactions on Image Processing* 30 (2021), pp. 9099–9111. DOI: 10.1109/TIP.2021.3118953. URL: https://arxiv.org/abs/2012.01724.