NHL STENDEN
computer vision & data science

Semantic Segmentation for Traffic Footage Monitoring

NHL Stenden Lectoraat in Computer Vision & Data Science

Diego Octavio Perez Urbina

Supervisor: Maya Aghaei Gavari

Abstract— The Netherland's cyclist culture has highlighted the increasing need for advanced traffic monitoring systems to ensure cyclist safety, especially given the 20% rise in bike-related fatalities between 2015 and 2023. Current methods rely mostly on manual video annotation, which is a time-consuming process and prone to human bias. This research proposes a framework for automating semantic segmentation in traffic camera footage, comparing the performance of two models on different image resolutions.
The research uses the Mapillary Vistas dataset for training and validation. To test the generalizability of the models, two frames from real-world footage were used. Results demonstrate the Mask2Former model with a Swin backbone ouperforms UNet across all evaluated metrics, achieving a 0.61 F1 Score in the validation split and 0.73 on an image from the Harlingen dataset.
These findings establish Mask2Former as a promising solution for reducing manual efforts in traffic monitoring across different footage. Future work will focus on fine-tuning models for grayscale datasets, testing UNet on higher resolution images, and exploring alternatives backbones for Mask2Former to improve its performance.

Index Terms—Semantic Segmentation, Mask2Former, UNet, Deep Learning in Traffic Analysis

1 Introduction

The Netherlands is well known for its vibrant cycling culture. This is well proven by the Netherlands Institute for Transport Policy Analysis, which stated that in 2016 there were a total of 4.5 billion bicycle trips [1], spanning a distance of 15.5 billion bicycle kilometers, and that a quarter of all trips made by Dutch residents are traveled by bicycle [2]. Therefore, the more bicycle trips the more common bike accidents. As stated by the Centraal Bureau voor de Statistiek, between 2015 and 2023, bike-related fatalities rose by 20%, accentuating the urgent need for better safety measures to protect cyclists. It is essential to implement better safety measures to sustain the increasing adoption of cycling.

Current traffic monitoring solutions rely on humans spending hours reviewing camera footage to label potentially dangerous behavior, and the first step to identify dangerous behavior is to identify interest areas. This process is time-consuming and prone to error and human bias. The development of automated tools for segmenting whole camera's scenes is vital for a sustainable way of monitoring bike traffic. These algorithms are designed to significantly reduce the time and effort humans spend analyzing traffic footage without dispensing human work.

This paper proposes a robust framework that automates the semantic segmentation of pixels from traffic camera footage. This will be accomplished by comparing various semantic segmentation models. Specifically, we aim to determine the most effective model for segmenting traffic scenes and assess its adaptability to datasets with varying characteristics, including grayscale and RGB data

1.1 State of the Art

In this section, we explored researchs using modified UNet structures for semantic segmentation

- *Diego Octavio Perez Urbina is an student from the AI For Image Data Minor at the NHL Stenden University of Applied Sciences, E-mail: diego.octavio.perez.urbina@nhlstenden.nl*
- *Maya Aghaei Gavari is a researcher lecturer in Computer Vision at the NHL Stenden University of Applied Sciences & Data Science, E-mail: maya.aghaei.gavari@nhlstenden.com.*

1.1.1 Cyclists Safety Factors

This research by PO Onaiyekan [3] concluded that cyclist-only accidents are the most prevalent form of cycling accidents, and more accidents with adverse outcomes were seen amongst elderly cyclists. Intersections are the most dangerous places on the road for cyclists and roundabouts are the safest because of the separation between cars and bikes[3]. Infrastructure-related risk factors could include potholes, wet fallen leaves, and wet cobblestones. Therefore the importance of identifying infrastructure on the footage.

1.1.2 Eff-UNet

In contrast to the original UNet [4] structure, Eff-UNet [5] uses an Efficient Net B7 as a decoder with pre-trained weights, and the original UNet as an encoder. The Efficient Net comprises 7 blocks of multiple Mobile Bottleneck Convolution layers and a fixed scaling factor set. The model was trained and tested with 320 x 224 resolution images from a subsample of India Driving Lite Dataset, which has urban and rural scenarios with complex obstructions and ambiguous road boundaries. The dataset has 7 distinct target classes, and data augmentation was used. After the training with 1404 images on 4-image batches and 10 epochs, the model reached a 0.73 mIoU on validation and 0.62 on testing, which was higher than other encoders combined with UNet decoding, and also outperformed earlier state of the art Convolutional Neural Networks(CNNs) regardless of being 8.4x smaller and 6.1x faster

1.1.3 MH-UNet Architecture

This article by M. Krithika [6] examines the evolution of UNet architectures for segmenting the region of interest from different internal organs. It proposes multiple modified UNet models from where MH-UNet [6] architecture was the best for tumor diagnosis with a Dice Score (DSC) of 0.90. The multi-scale UNet architecture consists of dense blocks, residual inception blocks, and hierarchical blocks that are included in the decoder and encoder. By doing so, it reduces the number of parameters and therefore the computational cost. The hierarchical blocks, which use dilated convolutional layers, are fundamental to capture multi-scale features. This approach capture spatial features of the image, which could help us a lot when segmenting spaces in the footage.

1.1.4 Asymm UNet for Crop Segmentation

UNet Architectures are known for their effectiveness in remote-sensing segmentation tasks, however, their computational cost is high because of the number of parameters. To mitigate this,

Asymm-UNet introduced Asymmetric Convolutions, kernels with different dimensions for height and width rather than using square kernels [7]. This modified architecture beats state-of-the-art models and it's focused on capturing spatial details and orientation-specific features, which is particularly relevant for monitoring abnormal behaviors

1.1.5 Semantic Segmentation With Attentional Atrous Feature Pooling

This [8] state-of-the-art CNN-based semantic object segmentation model consists of three stages: In the first stage, the feature map taken as input is passed through the maximum and average pooling blocks three times to create different sizes and obtain important features. In the second step, these combined features are multiplied by one feature obtained from them through the sigmoid block, and finally upsampling is applied to obtain the original input size. This method had an average IoU value of 70.59% which is 2% higher than previous state-of-the-art network models and could be used together with a UNet model to automatically mask bike lanes and then look for abnormal behavior in the lanes.

2 Materials and Methods

2.1 Datasets

For UNet model training, a subset of the Mapillary Vistas Dataset was used. The training split consists of 18,000 RGB images in PNG format with a 2560x1920 resolution and 8 bits per color channel, meaning each pixel value is between 0 and 255. Labels from the training split are in a PNG format, and in a 2D shape, where each pixel of the image represent a class id. Mask2Former training used the original Mapillary Vistas Dataset, which consists on 20,000 training images and 2,000 validation images with the same characteristics as the subset mentioned previously.

The class wise distribution of the training dataset can be found in Table 4. There are 18,000 images, 1,179,648,000 pixels and 66 distinct classes in the training split

The validation split consists on 2,000 images and labels with the same characteristic as the training split. You can see an example of an image and label from Mapillary dataset on Figure 1

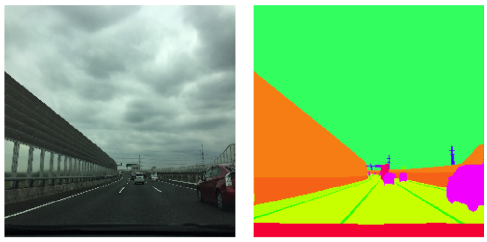

Fig. 1: Example of original image and its label from Mapillary Vistas

2.2 Dataset Pre-processing

For the UNet model training, several pre-processing steps were applied. Each image was resized to 256 × 256 resolution to ensure compatibility the model's architecture and to reduce training time and memory used. Pixel values were normalized to a range of 0 to 1 using the equation 1

$$\mathbf{I}_{\text{new}} = \frac{\text{image} - \mu_{\text{image}}}{\sigma_{\text{image}}} \tag{1}$$

Additionally, the images were converted to tensors, with their axes inverted to match the input requirements of the model.

In regards to the labels, they were converted from RGB images to 2d images where each pixel represents a class. This was acomplish by creating a hash table with each one of the unique RGB values and their respective class index. The color-class relationship was extracted from the .config file [9] found on the dataset original source.

2.3 Approaches

This research evaluates using 2 different models: UNet and Mas2Former with pre-trained weights

2.3.1 U-Net

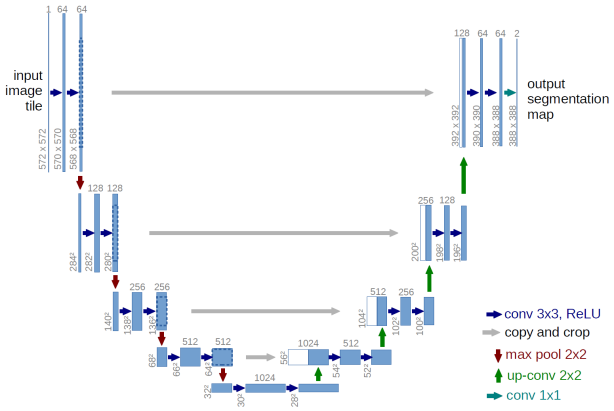

Fig. 2: UNet Architechture

UNet [4] consists of a contracting path and an expansive path. The contracting path applies two 3×3 unpadded convolutions, each followed by a rectified linear unit (ReLU), and a 2×2 max pooling operation with stride 2 for downsampling. At each downsampling step, the number of feature channels is doubled. The outputs from the encoder and decoder are concatenated along the channel dimension to facilitate skip connections, enhancing the model's ability to retain spatial information. The model receives 3-channel images as input and classifies each pixel into one of 66 distinct classes. The UNet architecture has a depth of 5 and the first convolutional layer utilizes 66 filters.

2.3.2 Mask2Former

Mask2Former [10] is a model that addresses instance, semantic, and panoptic segmentation. It consists of a backbone to extract image features, a multi-scale deformable attention Transformer as pixel decoder, and a query-based mask prediction head.

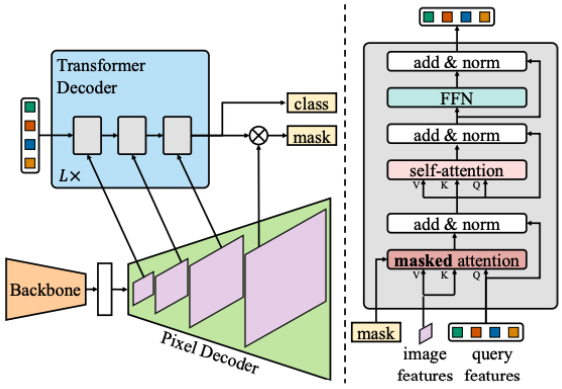

Fig. 3: Mask2Former Architechture

The backbone is responsible for extracting feature maps from the image to capture contextual information. The extracted features are then processed by the Transformer, which allows the model to

understand information across scales and spatial locations. Each mask query consists of a mask for each particular class, and the output is a combination of all class masks.

2.4 Metrics

To test the performance of both models, we employed the following metrics: Recall, Precission, and F1 Score.

Recall quantifies the model's ability to correctly identify all pixels belonging to a given class. A high recall value indicates that the model effectively detects the target class, minimizing missed pixels.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

Precision assesses the model's accuracy in predicting the pixels belonging to a specific class. High precision ensures that the predicted segmentation masks for a class contain minimal irrelevant pixels.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

F1 Score provides a balanced metric for evaluating the trade-off between precision and recall. It is the harmonic mean of precision and recall, capturing both the model's ability to correctly identify all relevant pixels and its accuracy in predicting positive pixels.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

3 Experiments and Results

3.1 Experiments

3.1.1 Mapillary Validation

The first experiment consisted on training an UNet model on Mapillary Vistas training subset, using an Adam optimizer with a learning rate of 0.001, 600 epochs, 18,000 training images, and 2,000 validation images from Mapillary validation split, with 256x256 resolution. Additionally, a learning rate scheduler was implemented, with patience 20, that reduces the learning rate when a plateau in validation loss is detected, helping to find local minimums. The model was evaluated on Mapillary Vistas validation split, because of the absense of a dedicated testing split. A Mask2Former [11] model from HuggingFace, pre-trained on the original Mapillary Vistas training split, using a 0.0001 Adam optimizer, 0.05 weight decay for all backbones , 16 batch size , and 50 epochs was also evaluated on Mapillary Vistas Validation split with different image resolutions. Evaluating both models on the Mapillary validation split was purely done to choose the model with the best performance.

3.1.2 Real World Footage

The model with the best performance was later evaluated on two real-world frames to understand how good the model generalize. The images were in JPEG format with a 1056x1056 resolution. To test the model accuracy The labels were manually generated for those 2 frames.

3.2 Results

When comparing the UNet model, and Mask2Former model with different resolution images, the results demonstrated that the best-performing model, Mask2Former with the Swin backbone, outperformed the UNet, and performed better on high-resolution images. The results can be found in the following Table: 1

Table 1: Comparison of model performance metrics on Mapillary validation split

Model	Resolution	Recall	Precision	F1 Score
Mask2Former (Swin Backbone)	1024x1024	0.63	0.60	0.61
Mask2Former (Swin Backbone)	512x512	0.62	0.60	0.60
Mask2Former (Swin Backbone)	256x256	0.60	0.59	0.58
UNet	256x256	0.33	0.48	0.35

The Mask2Former model trained on the higher-resolution images demonstrated considerable good performance when tested on the external images from Vinkeveen and Harlingen traffic camera footage. It demonstrated better performance on RGB images than with grayscale images.

Vinkeveen Example

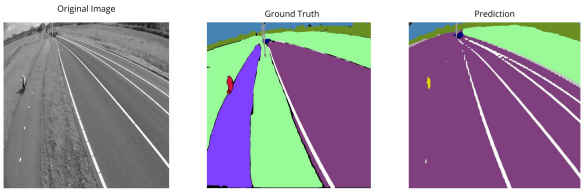

Fig. 4: Original image, ground truth, and prediction

Table 2: Vinkeveen Image Results

Image	Recall	Precision	F1 Score
Vinkeveen	0.56	0.66	0.54

Harlingen Example



Fig. 5: Original image, ground truth, and prediction

Table 3: Harlingen Image Results

Image	Recall	Precision	F1 Score
Harlingen	0.74	0.78	0.73

4 Discussion, Conclusion and Future Work

4.1 Discussion

This research serves as a preliminary study in the strengths and limitations of different models when applied to different datasets. On the Mapillary validation split, the Mask2Former with a Swin backbone outperformed the UNet model across all evaluated metrics. And within the Mask2Former models, the model tested in 1024x1024

resolution achieved the best results, demonstrating the importance of higher resolutions in enhancing accuracy.

Mask2Former demonstated promising adaptability. On the manually annotated Vinkeveen image, the model achieved a 0.54 F1 score, and conversly, it yielded a 0.73 F1 Score in the Harlingen image, suggesting that the model performs better in RGB Images. This result is consistent with the fact that the model was trained on RGB images. Additionaly, the superior performance at higher resolution aligns with the model's training on high-resolution images.

Something to highlight about the testing in the validation split, is that it could lead to overly optimistic performance scores because of data leakage, underscoring the necessity for an independent testing split.

4.2 Conclusion

This research highlights the effectiveness of a Mask2Former model with a Swin backbone for semantic segmentation in traffic monitoring solutions. The model outperformed UNet across all metrics, specially on a 1024x1024 resolution. Adittionaly, the model exhibited adaptability when tested on external datasets, specially on RGB images.

These findings highlight the potential of models like Mask2Former to significantly reduce manual labeling efforts in traffic monitoring systems. However, to improve the model's performance on datasets with characteristics different from those of Mapillary, fine-tuning will be necessary.

4.3 Future Work

While the results are promising, the model's lower performance on grayscale Vinkeveen dataset suggest a need for fine-tuning to adapt to grayscale images and data variations. Additionally, the model should be tested on more datasets, with a dedicated testing split from each one, to better understand its generalizability. These dataset should have labels, because of how time-consuming is to label high-resolution images.

In regards to the UNet model, future research should explore training UNet on higher-resolution images to understand whether its performance is resolution-dependent. Similarly, investigating Mask2Former model with alternative backbones may reveal additional performance improvements.

References

[1] Centraal Bureau voor de Statistiek. 684 road traffic deaths in 2023. https://www.cbs.nl/en-gb/news/2024/15/684-road-traffic-deaths-in-2023, 2024. Accessed: September 19, 2024.

[2] Ministerie van Infrastructuur. Cycling facts: New insights. https://english.kimnet.nl/publications/publications/ - 2020/11/03/cycling-facts-new-insights, 2023. Accessed: October 20, 2024.

[3] PO Onaiyekan, VL Passos, and K Horstman. A look at cycling safety in a southern municipality of the netherlands. *Cureus*, 14(5):e25268, May 24 2022.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[5] Bhakti Baheti, Shubham Innani, Suhas Gajre, and Sanjay Talbar. Eff-unet: A novel architecture for semantic segmentation in unstructured environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 358–359, 2020.

[6] M. Krithika alias AnbuDevi and K. Suganthi. Review of semantic segmentation of medical images using modified architectures of unet. *Diagnostics*, 12(12):3064, 2022.

[7] P. Goyal, A. Mitra, and M. Sinha. Asymm-unet: A concise model for segmentation of satellite imagery for crop mapping. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 9739–9743. IEEE, 2024.

[8] Jielin Jiang, Jiale Zhu, Muhammad Bilal, Yan Cui, Neeraj Kumar, Ruihan Dou, Feng Su, and Xiaolong Xu. Masked swin transformer unet for industrial anomaly detection. *IEEE Transactions on Industrial Informatics*, 19(2):2200–2209, 2022.

[9] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999, 2017.

[10] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation, 2022.

[11] Hugging Face Team and Meta AI. Mask2former - swin large mapillary vistas semantic. https://huggingface.co/facebook/mask2former-swin-large-mapillary-vistas-semantic, 2024. Accessed: 2024-12-27.

Supplementary Materials

Class-wise Distribution

Table 4: Class-wise Distribution of Images and Pixels

Class	Images	Total Pixels
Sky	17,803	345,449,560
Road	17,775	219,971,818
Pole	17,649	9,774,977
Unlabeled	17,546	20,269,816
Vegetation	17,471	178,247,980
Car	16,827	41,243,609
Building	16,478	146,366,290
Traffic Sign (Front)	16,297	5,293,682
Lane Marking - General	15,945	16,297,198
Curb	15,134	9,592,136
Sidewalk	13,143	34,544,875
Billboard	13,031	7,324,924
Fence	11,269	15,344,193
Utility Pole	9,167	4,908,317
Traffic Light	9,049	2,234,825
Person	8,985	3,686,568
Terrain	8,573	12,891,201
Wall	8,351	9,456,637
Lane Marking - Crosswalk	5,804	8,808,715
Ego Vehicle	5,446	23,455,452
Truck	4,339	4,446,249
Barrier	3,611	5,582,827
Guard Rail	2,993	3,263,825
Bus	2,887	3,267,745
Bridge	2,490	9,409,358
Parking	2,407	2,834,207
Mountain	2,282	2,606,865
Crosswalk - Plain	2,166	2,847,332
Bike Lane	1,334	4,143,392
Car Mount	1,226	1,464,719
Service Lane	1,157	2,157,578
Snow	652	3,930,495
Rail Track	640	1,614,831
Pedestrian Area	522	4,607,431