Deep Learning Semantic Segmentation for Road Scene Analysis

Long Zhuang, Wanshun Xu

Institute for Aerospace Studies

University of Toronto

Toronto, Canada

angela.zhuang@mail.com, wanshun.xu@mail.utoronto.ca

Abstract - This paper aims to present a comprehensive solution for road scene detection during driving. Motivated by the need for real-time capture of road events to ensure secure navigation, the study focuses on the enhancement of autonomous vehicle safety through the accurate identification of road elements. Leveraging the ENet architecture, a well-established deep learning model for pixel-wise semantic segmentation, the research introduces novel improvements, including the replacement of the activation function with Mish and the incorporation of Self-Attention Distillation(SAD). The proposed enhancements are systematically evaluated using the Cambridge-driving Labeled Video Database (CamVid), a popular dataset used to evaluate road objects classification. The comparative analysis demonstrates segmentation results in terms of Intersection over Union (IoU) and average IoU.

I. INTRODUCTION

Safety is a paramount concern in autonomous driving, with the real-time capture of road events being of utmost importance for secure navigation. Our project aims to address this critical need by focusing on obtaining comprehensive information about the drivable area. The overarching goal is to improve the safety of autonomous vehicles by preventing collisions and facilitating informed decision-making through the accurate identification of various road elements.

II. RELATED WORK

Traditional computer vision methods for road detection are typically categorized into shape and appearance models. The shape model employs Bezier Splines to represent road boundaries, utilizing RANSAC line fitting to provide initial estimates for a faster RANSAC algorithm. Optimal parameters for the bends are then determined through a randomly selected sample set [1]. In contrast, the appearance model, exemplified by [2], defines roads as a combination of diverse color maps, where each pixel's color distribution signifies its category. Traditional computer vision methodologies rely on external features such as color, shape, and topography within 2D images, refining road models through the assimilation of color cues across consecutive frames[3]. Nonetheless, these model-centric strategies face difficulties in congested road environments, particularly when visual attributes of the roads undergo substantial alterations.To address these challenges, traditional classifiers, including the structured random forest method, have been suggested for classifying image patches in a structured manner [4].

On the other hand, advanced deep learning methods, particularly deep learning-based semantic image segmentation architectures, have exhibited remarkable success in road detection. In recent years, Convolutional Neural Networks (CNNs) have emerged as powerful tools, demonstrating high feature extraction capabilities, accuracy, and efficiency [5]. Deep

learning models have proven to be particularly effective in solving semantic segmentation problems related to drivable area detection.

Various deep-learning models have been applied for pixel-wise semantic segmentation in, drivable area and lane detection. SegNet employs a combination of encoders and decoders to achieve pixel-wise semantic segmentation while optimizing memory usage and computational costs [6]. ENet, on the other hand, stands out for its high-speed encoder-decoder network, demonstrating superior performance with reduced computational demands compared to SegNet [7]. Spatial CNN (SCNN) introduces a novel approach to drivable area detection by implementing a message-passing procedure among consecutive pixels, enhancing segmentation performance at the expense of increased computational costs [8]. Additionally, other notable models like DeepLab [9] and Mask R-CNN [10] have found application in lane detection tasks. Pizzati et al. [11] propose a multi-task encoder-decoder network, further contributing to the diverse landscape of deep learning approaches for road scene understanding.

For the implementation of our deep learning model, we have strategically chosen ENet—an efficient neural network architecture renowned for its prowess in pixel-wise semantic segmentation. ENet distinguishes itself through its remarkable speed, demanding significantly fewer parameters and computational operations than alternative architectures, all while maintaining a high level of accuracy. This choice is pivotal in ensuring the scalability and real-time processing capabilities necessary for the dynamic environment of autonomous driving scenarios.

III. METHODOLOGY

A. ENet Model Architecture

The comprehensive architecture of the ENet network is detailed in Table 1, encompassing an initialization operation, an encoder-decoder structure, and a fully connected layer. Output sizes, provided for an illustrative input image resolution of 512 × 512, are presented in the table. The ENet network introduces a departure from the conventional encoder-decoder symmetry structure. Specifically, the convolution operation is streamlined in the decoder, resulting in enhanced efficiency and acceleration of image processing [27].

The ENet network commences with an initialization operation applied to the input image, as depicted in Figure 1a. Subsequently, the images undergo sequential processing through a MaxPooling layer with non-overlapping 2 × 2 windows and a convolution layer with 13 filters. The outcome of these operations, comprising 16 feature maps obtained through concatenation, serves the primary purpose of generating and fusing feature maps. This integration strategically combines the information derived from both pooling and convolution operations [27].

The ENet bottleneck module is primarily employed within the encoder-decoder structure, and its specific architecture is illustrated in Figure 2. Each bottleneck convolution module comprises three convolutional layers, arranged vertically as follows: a 1×1 projection map for dimensionality reduction, a primary convolutional layer, and a 1×1 expansion layer. The 'conv' operation within this module can take the form of a regular, dilated, or full convolution, or deconvolution, utilizing 3×3 filters, or a 5×5 convolution decomposed into two asymmetric ones [27]. Between these convolution layers, normalization and activation functions, such as PReLU or ReLU, are applied. This design facilitates the strategic integration of information derived from both pooling and convolution operations [28].

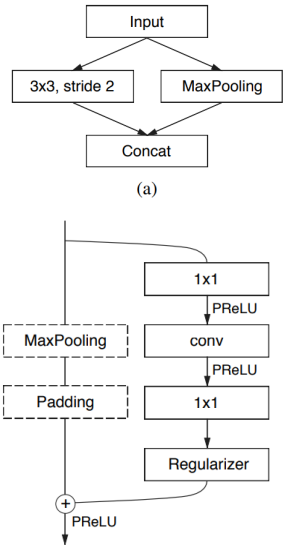


Figure 1: (a) ENet initial block. (b) ENet bottleneck module [27].

The ENet encoder-decoder section is organized into five parts. In the first part, a downsampled Bottleneck module is followed by four ordinary convolutional Bottleneck modules. The second part features two downsampled Bottleneck modules, followed by four Dilated Bottleneck modules, each with an expansion rate sequence of 2, 4, 8, 16, and 2 asymmetric Bottleneck modules within the stage. Both Stage 2 and Stage 3 share the same structure, with the exception that Stage 3 does not downsample the input initially. For Stages 4 and 5, tasked with returning the image to its original size, part 4 includes an upsampled Bottleneck followed by two regular Bottlenecks. Stage 5 consists of an upsampled Bottleneck and a regular Bottleneck. To incorporate regularization, Spatial Dropout is applied with a probability (p) of 0.01 before Bottleneck 2.0 and 0.1 afterward [27].

Table 1: ENet Architecture, Example input is given as 512 x 512 [27]

Name	Type	Output size
initial		$16 \times 256 \times 256$
bottleneck1.0	downsampling	$64 \times 128 \times 128$
$4\times$ bottleneck1.x		$64 \times 128 \times 128$
bottleneck2.0	downsampling	$128 \times 64 \times 64$
bottleneck2.1		$128 \times 64 \times 64$
bottleneck2.2	dilated 2	$128 \times 64 \times 64$
bottleneck2.3	asymmetric 5	$128 \times 64 \times 64$
bottleneck2.4	dilated 4	$128 \times 64 \times 64$
bottleneck2.5		$128 \times 64 \times 64$
bottleneck2.6	dilated 8	$128 \times 64 \times 64$
bottleneck2.7	asymmetric 5	$128 \times 64 \times 64$
bottleneck2.8	dilated 16	$128 \times 64 \times 64$
Repeat section 2, without bottleneck2.0		
bottleneck4.0	upsampling	$64 \times 128 \times 128$
bottleneck4.1		$64 \times 128 \times 128$
bottleneck4.2		$64 \times 128 \times 128$
bottleneck5.0	upsampling	$16 \times 256 \times 256$
bottleneck5.1		$16 \times 256 \times 256$
fullconv		$C \times 512 \times 512$

The overall architecture of ENet employs a strategy of downsampling by saving indices during max pooling layers. These saved indices are subsequently utilized in the decoder to produce sparse upsampled maps, contributing to the preservation of spatial details and mitigating common upsampling artifacts [28].

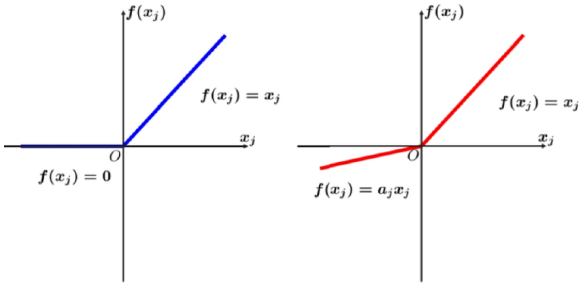
Early downsampling is strategically implemented in ENet to optimize the initial stages of the network and mitigate the computational costs associated with processing large input frames. Specifically, the first two blocks of ENet play a pivotal role in this optimization, as they significantly reduce the input size and leverage only a compact set of feature maps. This efficient downsampling mechanism not only enhances computational efficiency but also facilitates the network's ability to capture salient features in the early stages of processing, ultimately contributing to the overall effectiveness of ENet in real-time semantic segmentation tasks

B. Activation Function replaced with Mish

Mish is a novel self-regularized non-monotonic activation function, mathematically defined as $f(x) = x \cdot tanh(softplus(x))$ [29]

This activation function introduces a self-regularization mechanism, leveraging the soft plus and hyperbolic tangent ($tanh$) functions to enhance its non-linear properties [29]. The Mish activation function has demonstrated promising performance in various

deep learning applications, offering advantages in terms of training stability and convergence.



(a) ReLU Function **(b)** PReLU Function

Figure 2: (a) ReLU Activation Function (b) PReLU Activation Function

To understand the performance of the original ENet activation function, it is crucial to explore the functionalities of ReLU and PReLU. In Figure 2a, the graph of ReLU exhibits a switch-like behavior, distinguishing between the 'ON' and 'OFF' states, where weights greater than 0 are retained, and weights less than 0 are set to 0. While this mechanism addresses the vanishing gradients issue, it introduces the challenge of the dying ReLU problem, leading to inactive neurons. To tackle this challenge, the original ENet adopted PReLU, as depicted in Figure 2b. PReLU generalizes the traditional rectified unit by introducing a slope for negative values. Unlike ReLU, PReLU can learn the optimal slope for each neuron, effectively overcoming the dying ReLU problem [22].

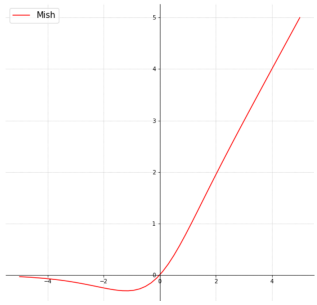


Figure 3: Mish Activation Function [22]

Figure 3 illustrates the graph of the Mish activation function, showcasing its unique characteristics. Notably, Mish effectively addresses the 'death ReLU' problem by retaining a relatively small negative value for weights, thus overcoming issues associated with the ReLU activation function. This approach allows Mish to mitigate the impact of large negative values on the system, preventing potential difficulties in learning specific patterns and enhancing the model's ability to capture features in images. The Mish activation function is designed with a smooth transition for negative values, helping to overcome saturation issues commonly associated with certain activation functions. By providing continuous gradients across its entire range, Mish promotes smoother learning dynamics. Its non-monotonic nature contributes to better learning by avoiding abrupt changes in activation, facilitating improved convergence during training.

Overall, Mish stands out as a versatile activation function, addressing challenges encountered by other functions and contributing to enhanced performance in image segmentation.

C. Self-Attention Distillation

To enhance road event segmentation accuracy, we incorporated Self-Attention Distillation (SAD) into the existing architecture, inspired by the work of Hou et al. [20] on lane detection. The objective is to improve the segmentation results by considering the interdependencies and importance of features across different layers.

1) Activation-based Self-Distillation

Consider the activation output of the m-th layer of the network as $A_m \in R^{C_m \times H_m \times W_m}$, where C_m, H_m, W_m

denote the channel, height, and width, respectively.An attention map is generated by applying a mapping function.

$$G^2_{sum}(A_m) = \sum_{i=1}^{C_m} |A_{mi}|^2$$

where A_{mi} is the i-th slice of A_m in the channel dimension [21]. This choice aligns with the mapping function used by Hou et al. for lane detection [20].

2) Integration into the Model

Attention generator, is represented by the function:

$$\Psi(\cdot) = \Phi(B(G^2_{sum}))$$

where $\Phi(\cdot)$ is spatial softmax operation, is added after each E2 and E3 encoder block of the ENet model.

3) Loss Formulation

Finally, distillation loss as illustrated below is added to the total loss starting at 70 epochs.

$$L = L_{seg}(s, \hat{s}) + \gamma L_{distill}(A_m + A_{m+1})$$

Where:

- L_{seg} is the standard cross-entropy segmentation loss.
- γ is a hyperparameter, initialized to 0.1, controlling the strength of the distillation loss.
- $L_{distill}(A_m + A_{m+1})$represents the distillation loss between the activation outputs$A_m$andA_{m+1}

This formulation combines the segmentation loss with the distillation loss, leveraging self-attention distillation for an objective to improve road event segmentation performance.

IV. EXPERIMENTAL. RESULTS

A. Dataset

The Cambridge-driving Labeled Video Database (CamVid) holds the distinction of being the inaugural collection of videos enriched with object class semantic labels and detailed metadata. This database offers ground truth labels that meticulously associate each pixel with one of 32 manually annotated semantic classes. These classes encompass a broad spectrum, including void, building, wall, tree, vegetation, fence, sidewalk, parking block, column/pole, traffic cone, bridge, sign, miscellaneous text, traffic light, sky, tunnel, archway, road, road shoulder, lane markings (driving), lane markings (non-driving), animal, pedestrian, child, cart luggage, bicyclist, motorcycle, car, SUV/pickup/truck, truck/bus, train, and other moving objects [25][26].

Originally captured as five video sequences, each featuring a 960×720 resolution camera mounted on the dashboard of a car, the database offers over ten minutes of high-quality footage at a frame rate of 30Hz [25][26]. Corresponding semantically labeled images are provided at 1Hz and, in part, at 15Hz, making CamVid an invaluable resource for road/driving scene understanding and an essential benchmark for semantic segmentation and object detection tasks [25][26].

B. Evaluation Metric

Intersection over Union (IoU) stands as a pivotal metric in image segmentation, particularly in semantic segmentation tasks. It quantifies the degree of overlap between the predicted and ground truth regions by computing the ratio of their shared area to their combined area [24]. IoU serves as a gauge for evaluating how accurately a model's predictions match the genuine objects within an image. A higher IoU value signifies improved precision in delineating the authentic boundaries and shapes of segmented objects, making it a key indicator of segmentation model performance. Written as a function of the confusion matrix between Actual(Y) and Prediction($\hat{Y}$)

	$\hat{Y}$=1	$\hat{Y}$=0
Y=1	TP	FN
Y=0	FP	TN

Where, TP=True positives, FP=False positives, etc., IoU is:

$$IoU(Y,\widehat{Y}) = \frac{TP}{TP+FN+FP}$$ [24]

The mIoU is then the average IoU across all classes:

$$mIoU(Y,\widehat{Y}) = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FN_i+FP_i}$$ [23]

C. Simulation Environment

The evaluation of ENet's performance was conducted on laptop GPUs equipped with NVIDIA GTX 3060. The assessment utilized the CamVid dataset, specifically emphasizing road scenarios, to showcase its real-time capabilities and accuracy in practical applications. This paper presents a comparative analysis between ENet and ENet+Mish+SAD, considering evaluation metrics such as Intersection over Union (IoU) and average IoU. This comparison offers a comprehensive insight into the segmentation results.

D. Neural Network Training

- Learning-rate = 5e-4
- Weight_decay = 2e-4
- Epoch = 200
- Criterion = CrossEntropyLoss Function
- Optimizer = Adam

E. Results and Analysis

The developed Encoder-Decoder system performs segmentation on the input image, assigning labels to individual pixels and generating a final output mask with color-coded representations of each object class. Given the intricate nature of urban surfaces, comprehensive object classification was essential to enable safe navigation for moving vehicles, ensuring avoidance of all obstacles.

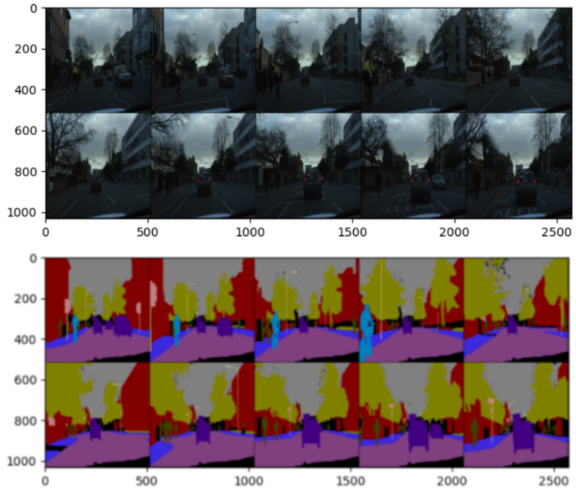


Figure 4: Test Dataset Input Image (shown at top) Ground Truth (shown at bottom)

In Figure 4, the Neural Network input and corresponding ground truth are illustrated. The system leverages the information provided by the ground truth to train and generate its segmentation output.

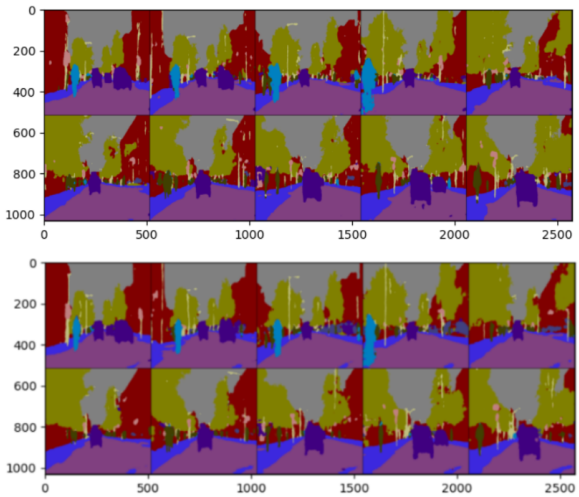


Figure 5: Prediction by original Enet (top) Prediction by Enet+mish+SAD (bottom

Examining Figure 5 reveals the outcomes of both the basic ENet and the Enhanced ENet (ENet+mish+SAD) in the context of the group test. Upon scrutiny, it is evident that both methods excel in picture segmentation. While the distinctions between the two are subtle, the Enhanced ENet exhibits smoother lines with fewer artifacts compared to its basic counterpart. Notably, in the original drawing (Figure 4), a

black area (unlabeled) below the ground truth is discernible, primarily labeled as the road surface (purple) by the basic ENet—a tendency common in deceiving interpretations. This occurrence is mirrored by the Enhanced ENet, but with an effort to categorize it as a distinct class other than the purple (road surface), aligning more accurately with its environmental context. This pattern extends to other unannotated black regions.

Moreover, the Enhanced ENet demonstrates a notable reduction in artifacts, resulting in greater continuity compared to the basic ENet. This trend suggests an improvement in segmentation quality with the Enhanced ENet, reinforcing its efficacy in achieving smoother and more accurate delineations of objects in the images.

Table 2 provides the IoU values for each labeled object, offering insights into the discrimination capabilities of each class. Notably, both methods exhibit commendable performance on larger objects, with a particular emphasis on roads—the primary label of interest. Enhanced ENet stands out with superior performance across various classes, including Sky, Building, Pole, Road, Tree, Pedestrian, and Bicycle.

However, challenges arise in identifying smaller objects such as Symbol and Fence, possibly attributed to inherent issues within the ENet architecture. The asymmetry in the encoding and decoding processes may contribute to these difficulties. Unlike UNet, ENet lacks the ability to preserve feature extraction during the encoding process, resulting in suboptimal performance in boundary areas. Despite these limitations, the overall performance, especially on significant objects like roads, remains promising and highlights the effectiveness of the improved ENet model.

Table 2: Results on CamVid test set (IoU value for each class)

Model	Sky	Building	Pole	Road	Pavement	Tree	Symbol	Fence	Car	Pedestrian	Bicyclist
ENet	89.92	65.25	20.72	91.92	**74.56**	61.48	**17.50**	**19.42**	**69.43**	30.60	33.69
ENet +Mish+SAD	**90.24**	**67.71**	**21.60**	**91.94**	72.27	**63.70**	17.09	16.33	65.68	**31.95**	**34.45**

Table 3: Results on CamVid test set (Final Testing loss and MIoU)

Modle	Final Testing loss	Mean IoU
ENet	1.0129	0.5223
ENet+Mish+SAD	1.0767	0.5209

Table 3 presents the Mean Intersection over Union (MIoU) and Final Testing loss for both models on the CamVid test set. While the MIoU values are approximately 52%, significant differences are observed in the Final Testing loss. Surprisingly, the group notes an unexpected result: the Enhanced ENet performs worse than the Basic ENet. The consensus within the group suggests that this unexpected outcome may be attributed to the absence of fine-tuning in the Enhanced ENet model.

Furthermore, the impact of unlabelled areas is considered significant in the performance of the Enhanced ENet. The model exhibits the ability to identify unlabelled regions, and it is hypothesized that these areas may contribute to the observed differences in performance metrics. This insight underscores the importance of considering fine-tuning and the handling of unlabelled areas when

evaluating and comparing the performance of segmentation models.

V. CONCLUSION

This report outlines an approach to enhance the road identification capabilities of intelligent cars, primarily leveraging encoder-decoder principles in deep learning. Following extensive research, the group designed the basic ENet, tested its performance on the CamVid Dataset, and observed satisfactory results. Subsequently, enhancements were introduced to the ENet model, involving the replacement of the encoder's activation function with Mish and an increased emphasis on Self-Attention Distillation. Despite marginal improvements in MIoU results, a visual comparison of generated images led the team to conclude that the enhanced ENet exhibits superior performance. The team believes that this improved model has the potential to significantly aid smart cars in recognizing roads and comprehending the intricacies of their driving environment.

REFERENCE

[1] M. Aly, "Real time detection of lane markers in urban streets", 2008 IEEE Intelligent Vehicles Symposium, 2008.

[2] Z. Chen and Z. Chen, "RBNet: A Deep Neural Network for Unified Road and Road Boundary Detection", Neural Information Processing Lecture Notes in Computer Science, pp. 677-687, 2017.

[3] Y. Gao, Y. Song and Z. Yang, "A real-time drivable road detection algorithm in urban traffic environment", International Conference on Computer Vision and Graphics, pp. 387-396, 2012, September.

[4] L. Xiao, B. Dai, D. Liu, D. Zhao and T. Wu, "Monocular Road Detection Using Structured Random Forest", International Journal of Advanced Robotic Systems, vol. 13, no. 3, pp. 101, 2016.

[5] G. L. Oliveira, W. Burgard and T. Brox, "Efficient deep models for monocular road segmentation", 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016.

[6] V. Badrinarayanan, 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, pp. 2481-2495, 2017.

[7] E. Romera, J. M. Alvarez, L. M. Bergasa and R. Arroyo, "ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation", IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 1, pp. 263-272, 2018.

[8] Y. Hou, Z. Ma, C. Liu and C. Change Loy, "Learning Lightweight Lane Detection CNNs by Self Attention Distillation", International Conference on Computer Vision (ICCV), 2019.

[9] Pizzati, F., Allodi, M., Barrera, A., & García, F. (2020). Lane detection and classification using cascaded CNNs. In Computer Aided Systems Theory–EUROCAST 2019: 17th International Conference, Las Palmas de Gran Canaria, Spain, February 17–22, 2019, Revised Selected Papers, Part II 17 (pp. 95-103). Springer International Publishing.

[10] Kumar, B., Garg, U., Prakashchandra, M. S., Mishra, A., Dey, S., Gupta, A., & Vyas, O. P. (2022, November). Efficient Real-time Traffic Management and Control for Autonomous Vehicle in Hazy Environment using Deep Learning Technique. In 2022 IEEE 19th India Council International Conference (INDICON) (pp. 1-7). IEEE.

[11] F. Pizzati and F. García, "Enhanced free space detection in multiple lanes based on single CNN with scene identification", 2019 IEEE Intelligent Vehicles Symposium (IV), pp. 2536-2541, 2019, June.

[20] Y. Hou, Z. Ma, C. Liu, and C. C. Loy, “Learning lightweight lane detection CNNS by self attention distillation,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019. doi:10.1109/iccv.2019.00110

[21] S. Zagoruyko and N. Komodakis. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In International Conference on Learning Representations, 2017.

[22] S. Rakshit, “SOUMIK12345/enet: Pytorch implementation of enet: A deep neural network architecture for real-time semantic segmentation (https://arxiv.org/abs/1606.02147),” GitHub, https://github.com/soumik12345/Enet (accessed Dec. 15, 2023).

[23] C. Nitr, “Miou calculation,” Medium, https://medium.com/@cyborg.team.nitr/miou-calculation-4875f918f4cb (accessed Dec. 15, 2023).

[24] C. Robinson, Understanding intersection-over-union,

https://calebrob.com/ml/2018/09/11/understanding-iou.html (accessed Dec. 15, 2023).

[25] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from Motion Point Clouds," Lecture Notes in Computer Science, pp. 44–57, 2008. doi:10.1007/978-3-540-88682-2_5

[26] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," Pattern Recognition Letters, vol. 30, no. 2, pp. 88–97, 2009. doi:10.1016/j.patrec.2008.04.005

[27] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation,"arXiv.org, https://arxiv.org/abs/1606.02147 (accessed Dec. 15, 2023).

[28] Y. Wang, "Remote Sensing Image Semantic segmentation algorithm based on improved Enet Network," Scientific Programming, vol. 2021, pp. 1–10, 2021. doi:10.1155/2021/5078731

[29] D. Misra, "Mish: A self-regularized non-monotonic activation function," arXiv.org, https://arxiv.org/abs/1908.08681 (accessed Dec. 15, 2023).