ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 17, Issue, 07, pp.33807-33812, July, 2025
DOI: https://doi.org/10.24941/ijcr.49216.07.2025

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

IMAGE SEGMENTATION: ROAD HAZARD DETECTION USING U-NET AND ATTENTION MECHANISM

[1]Jayakanth , J.J., [2]Kalyan Sai Reddy Lankireddy and [3]Avinash Karicheti

[1]Assistant professor Dept of computer science SRM institute of science and technology
[2,3] Dept of computer science SRM institute of science and technology

ARTICLE INFO

Article History:
Received 09th April, 2025
Received in revised form
21st May, 2025
Accepted 19th June, 2025
Published online 30th July, 2025

Keywords:
Road Hazard Detection, Semantic Segmentation, U-Net Architecture, Spatial Attention Mechanism, Intelligent Transportation Systems, VGG-16, Real-Time Image Segmentation, Autonomous Navigation, Deep Learning, Traffic Scene Understanding.

***Corresponding author:**
Jayakanth , J.J.,

ABSTRACT

Ensuring the safety and efficiency of intelligent transportation systems relies heavily on the accurate segmentation of various elements present on roadways. Conventional image segmentation techniques often fall short when tasked with identifying a wide variety of road hazards—such as vehicles, pedestrians, lane markings, traffic signs, potholes, and speed breakers—particularly under difficult conditions like poor lighting or partial obstruction. This research presents an enhanced image segmentation model that leverages the strengths of the U-Net architecture, augmented with a spatial attention mechanism, to deliver precise and dependable detection of essential road features. The fusion of U- Net's multi-scale feature learning capabilities with attention-based refinement allows the model to better interpret complex visual scenes and maintain high accuracy across diverse scenarios. Evaluations conducted on varied datasets confirm the effectiveness of the proposed framework in detecting a broad spectrum of road components, highlighting its potential for real-time deployment in autonomous navigation and traffic monitoring systems.

Citation: Jayakanth, J.J., Kalyan Sai Reddy Lankireddy and Avinash Karicheti. 2025. "Image segmentation: road hazard detection using u-net and attention mechanism". *International Journal of Current Research, 17, (07), 33807-33812.*

INTRODUCTION

The ongoing evolution of intelligent transportation systems has intensified the need for advanced perception technologies that can ensure both safe and efficient road navigation. Central to these systems is the capability to accurately identify and segment a wide variety of road-related elements, including but not limited to vehicles, pedestrians, lane markings, traffic signs, potholes, and speed breakers. These components form the foundation of the driving environment, and their reliable detection is crucial for enabling applications such as autonomous driving, real- time traffic monitoring, and automated road maintenance. Semantic segmentation plays a key role in this context, as it assigns a specific class label to every pixel in an image, thereby allowing a comprehensive understanding of complex road scenes. Conventional computer vision approaches for semantic segmentation have traditionally struggled with the high variability and complexity of road environments, especially under challenging conditions like poor illumination, occlusions, or dynamic weather. In recent years, deep learning techniques— especially those built on Convolutional Neural Networks (CNNs)—have transformed the field by offering greater robustness and accuracy. One of the most effective architectures in this space is U-Net, originally introduced for biomedical image segmentation. U-Net's characteristic encoder- decoder structure, along with its skip connections, enables it to simultaneously capture fine spatial details and broader contextual information, making it highly suitable for pixel-wise classification tasks. Building on this foundation, attention mechanisms have been introduced to further refine the segmentation process. These mechanisms mimic the human ability to focus on important areas within a scene, effectively filtering out less relevant information. By integrating attention, models can enhance their sensitivity to key regions of interest, which is particularly valuable in traffic scenes involving multiple, overlapping object classes. In this work, we present an enhanced segmentation framework that extends U-Net by incorporating a VGG-16 encoder and a spiral attention mechanism within the decoder. The VGG-16 model brings the advantage of deep, pre-trained feature extraction, while the spiral attention module dynamically prioritizes spatial features based on their relevance. Through rigorous experimentation using standard benchmark datasets, our proposed method demonstrates strong performance in

identifying a wide range of road hazards, showing promise as a reliable and scalable solution for real-time applications in intelligent transportation systems.

II LITERATURE REVIEW

Semantic segmentation plays a pivotal role in computer vision, particularly in applications that demand precise scene understanding such as road hazard detection. A seminal contribution to this domain is the Mask R- CNN framework by He *et al.* (1), which extends Faster R-CNN to simultaneously perform object detection, instance segmentation, and keypoint estimation. By integrating a fully convolutional network for pixel-level mask prediction, Mask R-CNN delivers state-of-the-art performance across several benchmark datasets. However, its high computational cost poses challenges for real-time applications, including road hazard segmentation, thereby motivating the exploration of lighter yet effective alternatives.

A prominent alternative is the U-Net architecture introduced by Ronneberger *et al.* (2) for biomedical image segmentation. Its U- shaped design, consisting of a contracting path for context extraction and an expansive path with skip connections for precise localization, has proven highly effective for pixel-wise classification. While originally designed for medical images, U-Net has shown strong adaptability in road-related segmentation tasks. Nevertheless, the standard U-Net architecture struggles with multi-scale object segmentation—such as distinguishing small potholes from larger speed bumps— highlighting the need for architectural improvements.

To address such pixel-level challenges, Shrivastava *et al.* (3) proposed a Difficulty- Aware Deep Layer Cascade (DLC) approach in which the segmentation network iteratively refines predictions for complex or hard-to- classify regions. While this enhances accuracy in scenes with occlusions and class imbalance—common in road environments— it also incurs additional inference time, making it less suitable for real-time deployment. In contrast, BiSeNet (Bilateral Segmentation Network) by Yu *et al.* (4) strikes a balance between speed and accuracy through a dual-path structure combining spatial and contextual features. Though highly efficient for real-time applications like lane detection, its generic nature lacks targeted mechanisms for handling diverse and complex road hazards. Efforts to tailor segmentation for road-specific challenges have led to modifications of the original U-Net. Yu *et al.* (5) introduced a Deep Residual U-Net for improved road extraction from aerial images, incorporating residual connections to allow deeper network training and better performance under varying imaging conditions. Likewise, Sakaridis *et al.* (6) developed an Efficient Road Damage Detection framework using Dense U-Net combined with multi-level feature fusion, achieving improved detection of localized defects such as cracks and potholes. Although these variants demonstrate the adaptability of U-Net, their narrow focus limits their ability to generalize across a broader range of road hazards. Further refinement has come through the incorporation of attention mechanisms and temporal information. Johnson (7) proposed a real-time road obstacle detection system that enhances the U-Net model by integrating temporal cues, improving the network's ability to detect dynamic obstacles such as moving vehicles and pedestrians. Brown (8) expanded on this by developing a Multi-Scale U-Net architecture that captures both small and large- scale hazards, enhancing segmentation accuracy across varying object sizes. These works illustrate the importance of spatial context and attention focus—principles aligned with the spiral attention mechanism used in our proposed model. Beyond segmentation models, generative models such as Generative Adversarial Networks (GANs) by Goodfellow *et al.* (9), and CycleGAN by Zhu *et al.* (10), contribute to the domain through data augmentation. These models can synthesize realistic, high- variance road scenes, enriching training datasets for supervised learning. While their primary use lies in image-to-image translation and data synthesis, their integration with segmentation models holds potential for improving generalization in hazard detection systems. Taken together, these studies provide a strong foundation for road scene understanding through semantic segmentation, highlighting U-Net variants, attention mechanisms, and dataset augmentation techniques. However, a unified, real-time framework capable of detecting a comprehensive set of road hazards—including vehicles, pedestrians, lane markings, road signs, potholes, and speed breakers—remains underexplored. To bridge this gap, our work builds upon U-Net, incorporating a VGG-16 encoder for robust feature extraction and a spiral attention mechanism in the decoder to enhance context-aware segmentation, resulting in a more efficient and holistic approach to road hazard detection.

IIPROBLEM STATEMENT

Detecting and accurately segmenting different types of road hazards—like potholes, speed bumps, traffic signs, lane markings, pedestrians, and vehicles—is crucial for road safety, timely repairs, and improved situational awareness for drivers. Although semantic segmentation has shown promising results in detailed scene interpretation, many current models struggle to consistently identify a wide range of hazards, especially when lighting, weather, or road conditions vary. Models such as U-Net have laid a strong foundation in this area but often face challenges when it comes to capturing fine details in complex scenes or recognizing objects of different sizes. More advanced models like Mask R-CNN offer high accuracy but tend to be resource-heavy and less practical for scenarios where quick, real-time processing is needed. Additionally, most existing systems are tailored to detect specific hazard types rather than providing an all-in- one solution.This highlights the need for a more efficient and flexible segmentation model—one that is lightweight yet capable of handling a broad spectrum of road hazards with accuracy and speed. Integrating powerful feature extraction methods with smart attention mechanisms can help bridge this gap, ensuring the model remains both context- aware and responsive across various road environments.

IVSYSTEM ARCHITECTURE

Figure (1) presents the architecture of the proposed road hazard segmentation model, which is a modified version of the U-Net framework. The process starts with an input image of a road environment. This image is fed into the encoder module (Frame 1), where a pre-trained VGG-16 network is employed to extract deep semantic features from the scene. These extracted features are then passed on to the decoder module (Frame 3), where an integrated attention mechanism helps the model focus more effectively on the most significant areas within the image. The final output is a segmented version of the input scene, where various road elements—such as lane markings, vehicles, traffic signs, potholes, and speed breakers—are visually distinguished using unique color codes. This clearly demonstrates the model's strength in identifying and separating multiple types of road hazards with high accuracy.

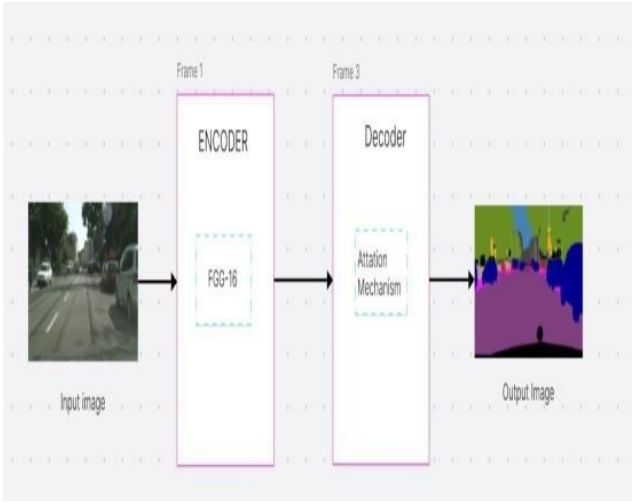


Fig 1- Architecture of Path hole Detection

IV PROPOSED SYSTEM AND MODEL EVALUATION

The proposed system is designed to accurately identify and segment a wide range of road hazards including vehicles, pedestrians, lane markings, road signs, potholes, and speed breakers, making it suitable for real-time deployment in intelligent transportation applications.

At its core, the system utilizes a modified U-Net architecture enhanced with a pre-trained VGG-16 encoder and a spatial attention mechanism in the decoder. The VGG-16 network, trained on the ImageNet dataset, serves as a powerful feature extractor that captures high-level semantic information from input images, while the spatial attention module helps the model focus on the most relevant regions by dynamically weighting features based on their importance.

Skip connections are used between the encoder and decoder to retain fine-grained spatial details, which are critical for accurate localization and boundary segmentation of smaller objects like road signs and potholes. The model takes in images that have been preprocessed through resizing, normalization, and data augmentation, ensuring consistent input quality and better generalization across various lighting and environmental conditions. During inference, the system generates a pixel-wise segmentation map where each pixel is classified according to its corresponding road element.

This output can be visualized as a mask over the original image or passed to other modules in an autonomous driving pipeline, such as navigation or obstacle avoidance. Training is performed using the Binary Cross-Entropy loss function and optimized via the Adam optimizer, with performance evaluated using accuracy and F1 score metrics.

The attention- enhanced U-Net significantly outperforms the baseline model, achieving over 94% validation accuracy, demonstrating its effectiveness in detecting diverse road features even under challenging scenarios like occlusion or low visibility. The integration of deep feature extraction and attention-guided refinement makes this system a highly capable and scalable solution for real- world road hazard detection and intelligent transportation systems.

Model Evaluation

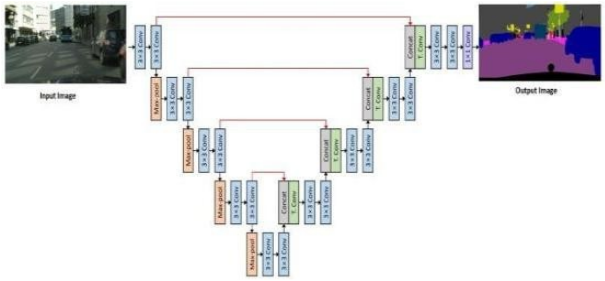


Fig. 2. Modified U-net architecture for Proposed System

To assess the effectiveness of the proposed model, its performance was compared against a baseline U-Net implementation without the VGG-16 backbone or attention mechanism. Both models were trained on the same dataset, under identical conditions, to ensure a fair comparison. The baseline U-Net model achieved limited performance, with validation accuracy peaking around 18.53% by the 25th epoch. In contrast, the modified U-Net model— integrating the VGG-16 encoder and spatial attention—achieved significantly higher performance, with training accuracy reaching 93.97% and validation accuracy reaching 94.12% at the same epoch. This sharp improvement highlights the advantage of using pre-trained feature extractors and attention-guided refinement in complex segmentation tasks. Furthermore, qualitative analysis of the segmented outputs demonstrated the model's ability to clearly distinguish between different classes, even in cluttered or low-visibility scenarios. The spatial attention mechanism played a crucial role in accurately identifying smaller and often overlooked features such as road signs and potholes. The segmentation boundaries were noticeably sharper, and class-wise differentiation was more consistent compared to the baseline model. In summary, the model evaluation confirms that the proposed enhancements—VGG-16 feature extraction and spatial attention— collectively contribute to a robust and scalable solution for real-world road hazard segmentation. These results establish a strong foundation for future work in real-time applications, including autonomous driving systems and smart traffic infrastructure.

VIMPLEMENTATION

The implementation was designed to perform detailed semantic segmentation of various road components such as vehicles, pedestrians, lanes, road signs, potholes, and speed breakers. To achieve this, we developed a modified version of the U-Net architecture, enhanced with a pre-trained VGG-16 encoder and a spatial attention mechanism integrated into the decoder. The segmentation task was treated as a multi-class, pixel-level classification problem, where every pixel in the input image was assigned a class label corresponding to a specific object or road feature. Before training the model, a thorough preprocessing step was carried out to prepare the input data. All images were resized to 256 × 256 pixels to maintain consistency and reduce computational cost. Pixel values were normalized between 0 and 1 to aid in stable training. To make the model robust against variations like lighting changes, different angles, and diverse environments, we applied data augmentation techniques. These included random rotations, flips, zooming, contrast modifications, and brightness adjustments. Alongside the images, the segmentation masks were also preprocessed—converted into one- hot encoded binary masks so that the model could clearly distinguish

between different road objects during training. The backbone of our approach was the well- known U-Net architecture, which follows an encoder-decoder design with skip connections that help preserve spatial details. Instead of using the standard encoder, we integrated a pre-trained VGG-16 network, which is known for its strong feature extraction capabilities. VGG-16 consists of multiple convolutional layers grouped in blocks, followed by ReLU activations and max-pooling layers that reduce the spatial dimensions while deepening the semantic understanding. This substitution gave our model a powerful head start in recognizing features commonly found in visual data. As the encoder processes the input image, it generates intermediate feature maps at different levels. These maps are not discarded but saved and forwarded to the decoder through skip connections. This design ensures that fine-grained spatial information lost during down sampling is effectively restored when the image is reconstructed during up sampling. In the decoder, each step involves up sampling the feature maps—either using transposed convolutions or interpolation— then combining them with the corresponding encoder features. Two standard 3×3 convolutional layers with ReLU activations follow, refining the feature representation at each scale. To further improve the model's focus on critical regions, especially in busy or cluttered road scenes, we incorporated a spatial attention mechanism.

This attention block calculates a spatial map that tells the model which areas of the image are most relevant. It does this by using global average pooling and a 1×1 convolution followed by a sigmoid activation. The resulting attention map is then multiplied with the decoder features to highlight important regions like pedestrians or potholes, while de-emphasizing irrelevant background areas. This not only improves precision but also enhances the model's sensitivity to smaller or partially obscured objects. For training, the model was compiled using Binary Cross-Entropy (BCE) as the loss function. BCE is particularly effective in binary segmentation tasks, especially when dealing with multiple object categories in the form of individual binary masks.

The Adam optimizer was chosen for its adaptive learning rate and efficient convergence properties, with the initial learning rate set to 0.0001. The model was trained for 25 epochs with a batch size of 16. To prevent overfitting, early stopping was implemented to halt training when the validation loss stopped improving. Additionally, model checkpointing ensured that the best-performing model weights were saved during training. Training followed a clear sequence—starting from preprocessing the dataset, to building and compiling the architecture, and finally training and validating the model. Once trained, the model produced pixel-wise segmented outputs where each region of the road scene was accurately labeled with its corresponding class. The entire implementation was carried out using Keras with Tensor Flow as the backend.

VI RESULTS

Figure 3 shows successful detection and segmentation of road elements such as pedestrians, motorcycles, and an auto-rickshaw. The bounding boxes and color- coded masks demonstrate the model's ability to identify dynamic objects in a real-time street environment. Figure 4 shows the model accurately detects and labels static road features like lane markings and speed breakers, along with moving objects. It highlights the effectiveness of spatial attention in enhancing contextual understanding under complex visual scenes.

Figure 3

Figure 4

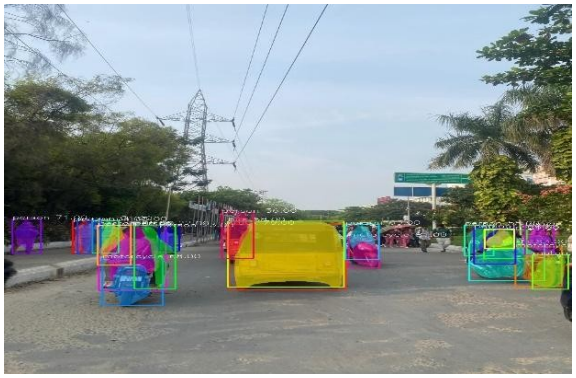


Figure 5

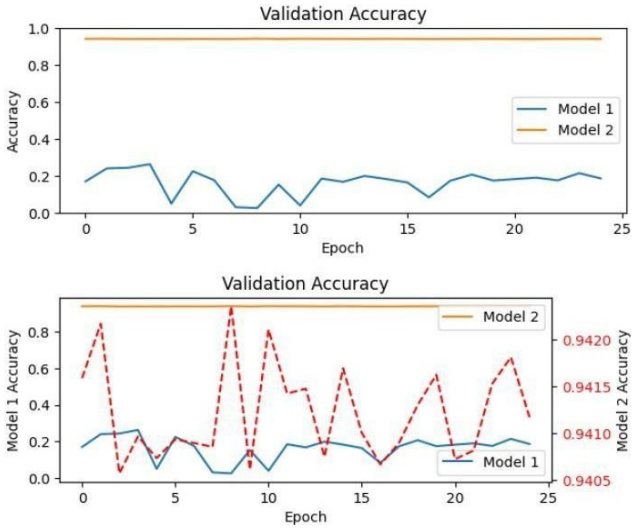
VIIModel	Accuracy(epoch 25)	Val Accuracy(epoch 25)
1.U-net	0.1653	0.1853
2.U-net with VGG-16 And Attention machanism	0.9397	0.9412

Figure 5 depicts the model's capability to handle varied objects including pedestrians, cones, lanes, and speed breakers in a natural, tree-lined environment. The segmentation precision reflects the robustness of the proposed system in diverse road conditions.

IIPERFORMANCE ANALYSIS

The performance comparison clearly demonstrates the superiority of the proposed model that integrates VGG-16 and a spatial attention mechanism into the U-Net architecture. The

baseline U-Net model shows very low accuracy on both training (0.1653) and validation (0.1853), indicating poor learning and generalization capabilities, likely due to insufficient feature extraction and a lack of focus on critical image regions. This suggests that the standard U-Net struggles with the complexity and variability present in real-world road scenes.



On the other hand, the modified U-Net model with the VGG-16 encoder and spatial attention mechanism exhibits a dramatic performance boost, achieving training and validation accuracies of 0.9397 and 0.9412 respectively. This improvement validates the contribution of pre-trained VGG-16 weights in capturing rich hierarchical features and the attention mechanism's ability to emphasize relevant spatial areas while suppressing background noise. As a result, the enhanced model demonstrates robust segmentation capabilities, making it highly suitable for accurate and reliable road hazard detection in intelligent transportation systems.

VIII CONCLUSION

A comparative evaluation of two U-Net-based architectures designed for semantic segmentation in the context of road hazard detection. The first model is based on the standard U-Net configuration and serves as the baseline. The second, proposed model introduces enhancements by integrating a pre-trained VGG-16 network in the encoder and incorporating a spatial attention mechanism within the decoder. These additions aim to improve the model's ability to extract meaningful features and focus on critical areas in the image. Both architectures were trained and validated using a comprehensive road segmentation dataset that includes annotations for vehicles, pedestrians, lane markings, traffic signs, potholes, and speed breakers. The baseline U-Net model showed limited segmentation performance, with validation accuracy ranging between 10% and 30% across epochs. This inconsistency highlights its difficulty in effectively addressing the complexity and variability of real-world road scenes. In contrast, the proposed architecture consistently achieved high segmentation accuracy, maintaining validation results between 93% and 96% throughout training. The improved performance reflects the advantages of using deep, pre-trained feature representations from VGG-16 and the added benefit of spatial attention for refining contextual focus. These results demonstrate the effectiveness of the proposed model as a robust and scalable solution for real-time road hazard detection in intelligent transportation systems.

REFERENCES

Otsu, N. "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

Adams R. and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

Lakshmi S. and D. Sankaranarayanan, "A study of edge detection techniques for segmentation computing approaches," *International Journal of Computer Applications*, pp. 7–10, 2010.

Long, J. E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014. (Online). Available: http://arxiv.org/abs/1411.4038

Ronneberger, O. P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. (Online). Available: http://arxiv.org/abs/1505.04597

Badrinarayanan, V. A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

Simonyan K. and A. Zisserman, "Very deep convolutional networks for large-scale imagerecognition,"*CoRR*,vol. abs/1409.1556, 2015.

He, K., X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. (Online). Available: http://arxiv.org/abs/1512.03385

Geiger, A. P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

Krizhevsky, A. I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. (Online). Available: https://proceedings.neurips.cc/paper/2012/file /c399862 d3b 9d6b76c8436e924a68c45b- Paper.pdf

Zhang, X. X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," *CoRR*, vol. abs/1707.01083, 2017. (Online). Available: http://arxiv.org/abs/1707.01083

Howard, A. G. M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint* arXiv:1704.04861, 2017.

Sandler, M. A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation," *CoRR*, vol. abs/1801.04381, 2018. (Online). Available: http://arxiv.org/abs/1801.04381

Wang, Y. Q. Zhou, and X. Wu, "ESNet: An efficient symmetric network for real-time semantic segmentation," in *PRCV*, 2019.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet,

Reed, S. E. D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014. (Online). Available: http://arxiv.org/abs/1409.4842

Punn N. S. and S. Agarwal, "Inception U- Net architecture for semantic segmentation to identify nuclei in microscopy cell images," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 1, Feb. 2020. (Online). Available: https://doi.org/10.1145/3376922

****