MATHEMATICAL MODELING AND COMPUTING, Vol. 10, No. 3, pp. 854–863 (2023)

Improving pedestrian segmentation using region proposal-based CNN semantic segmentation

Lahgazi M. J.[1], Argoul P.[2], Hakim A.[1]

[1]*Faculty of Sciences and Technics, Cadi Ayyad University, Marrakesh, Morocco*
[2]*MAST-EMGCU, Univ Gustave Eiffel, IFSTTAR, F-77477 Marne-la-Vallée, France*

(Received 26 March 2023; Accepted 18 May 2023)

Pedestrian segmentation is a critical task in computer vision, but it can be challenging for segmentation models to accurately classify pedestrians in images with challenging backgrounds and luminosity changes, as well as occlusions. This challenge is further compounded for compressed models that were designed to deal with the high computational demands of deep neural networks. To address these challenges, we propose a novel approach that integrates a region proposal-based framework into the segmentation process. To evaluate the performance of the proposed framework, we conduct experiments on the PASCAL VOC dataset, which presents challenging backgrounds. We use two different segmentation models, UNet and SqueezeUNet, to evaluate the impact of region proposals on segmentation performance. Our experiments show that the incorporation of region proposals significantly improves segmentation accuracy and reduces false positive pixels in the background, leading to better overall performance. Specifically, the SqueezeUNet model achieves a mean Intersection over Union (mIoU) of 0.682, which is a 12% improvement over the baseline SqueezeUNet model without region proposals. Similarly, the UNet model achieves a mIoU of 0.678, which is a 13% improvement over the baseline UNet model without region proposals.

Keywords: *pedestrian segmentation; region proposals; UNet; object detection; semantic segmentation; convolutional neural networks (CNNs).*

2010 MSC: 68T45, 68U10, 94A08 **DOI:** 10.23939/mmc2023.03.854

1. Introduction

With the rapid advancement in technology, there has been an exponential growth of image data in various real-world applications. Consequently, image understanding has become a crucial area of computer vision research, which involves the identification and localization of objects within an image [1]. One important task in this field is object detection which has been extensively studied. Traditional object detection methods are built on handcrafted features and shallow trainable architectures, but their performance easily stagnates [2–5]. With the development of deep learning, more powerful tools have been introduced to address these problems. Faster R-CNN is one such method that introduces region proposal networks (RPNs) to generate region proposals and performs object detection on the proposals [6]. You Only Look Once (YOLO) is another widely used object detection method that directly predicts bounding boxes and class probabilities from the whole image [7,8]. YOLO has gained popularity due to its high accuracy and fast inference speed. CornerNet is another object detection method that detects objects as paired keypoints [9].

A more straightforward task is image segmentation which has garnered significant interest in recent years [10]. Image segmentation is the process of partitioning an image into multiple segments or objects. Two primary types of image segmentation are semantic and instance segmentation. Semantic segmentation is a pixel classification problem with semantic labels. Its objective is to assign labels to pixels within an image by dividing it into different semantic regions. Instance segmentation is an enhanced form of semantic segmentation that involves the detection of instances for each category. Its aim is to assign a unique label to each instance of a class in an image [10].

In recent years, deep learning algorithms have made significant progress in image segmentation. Among these algorithms, Convolutional Neural Networks (CNNs) have emerged as one of the most effective and widely used architectures for image segmentation in the deep learning community [11–18]. This is primarily due to the fact that CNNs combine the capabilities of both feature extraction and classification, making them highly effective for this task. CNNs can learn complex feature representations directly from raw data, enabling them to capture the visual appearance and shape of objects. Consequently, CNNs have been highly successful in object segmentation, including pedestrian semantic segmentation, as evidenced by their strong performance in recent studies [16–18].

Interest in improving the accuracy of convolutional neural network (CNN) architectures for image segmentation has grown recently. However, the trend towards constructing very deep and complex neural networks that achieve high accuracy has resulted in a significant demand for memory and computational resources, limiting their applicability in real-world scenarios [19]. These resource requirements are especially problematic for devices with limited resources and restricted power and memory, such as online learning and mobile phones [20]. Therefore, reducing the storage requirement and computational cost of neural networks has become a crucial area of research. In response, modern CNNs have been designed to be more effective in real-world environments, and numerous studies have been conducted to address these challenges.

The growing demand for more efficient and faster convolutional neural networks (CNNs) has led to research efforts focused on reducing their storage and computational requirements. One popular approach is the use of pruning techniques, which aim to remove redundant parameters that do not affect network accuracy, thereby resulting in faster and more compact deep CNNs [21–26]. Additionally, factorization and low-rank approximation methods have been applied to exploit the linear structure of neural networks and reduce redundancy in deep learning models [27–32]. Furthermore, researchers have designed small network architectures such as SqueezeNet that utilize depthwise separable convolutions to reduce parameters while maintaining similar accuracy to larger networks [33–38]. For image segmentation tasks, compact filters have been proposed to replace over-parameterized filters in existing networks, leading to models like SqueezeUNet that balance efficiency and performance [39,40]. However, some of these solutions tend to compromise performance for efficiency.

In this paper, we propose a novel framework that leverages region proposal to enhance the accuracy of semantic segmentation in compressed models. Our approach addresses the challenge of preserving the performance of compact models while maintaining high segmentation accuracy. We utilize YOLO as a region proposal model and compare the performance of UNet and SqueezeUNet for semantic segmentation tasks with and without region proposals. Additionally, we utilize the OCHuman dataset to train the models and evaluate our proposed framework. The results demonstrate the effectiveness of our approach in achieving improved segmentation accuracy while maintaining the compression benefits of the models.

2. The proposed model: Region proposals based pedestrian segmentation framework

In this section, we provide an overview of the pipeline of our proposed framework, including the models and datasets used for evaluation. First, we introduce the Yolo model, which is employed for generating region proposals. Next, we describe two segmentation models used in our experiments, namely UNet [39] and SqueezeUNet [40], and explain how they are used to evaluate the performance of our proposed framework. Then, we explain the pipeline of the proposed framework. Finally, we introduce the OCHuman dataset, used to train both segmentation models.

YOLO (You Only Look Once)

YOLO (You Only Look Once) is a state-of-the-art object detection model that is widely used in computer vision research and applications. It is a neural network architecture that uses a single convolutional network to simultaneously predict object bounding boxes and class probabilities. Compared to previous versions, YOLOv3 has improved accuracy and speed, thanks to a variety of techniques including feature pyramid networks, improved training strategies, and darknet-53 as the backbone network. YOLOv3 can handle detection tasks on a wide range of objects and scenes, making it a popular choice for real-time applications such as self-driving cars, surveillance systems, and robotics. There are more recent versions of YOLO, but we used v3 as it is largely sufficient of our use case. In our work, we have adapted the YOLO v3 architecture to focus only on pedestrian detection, rather than identifying all classes of objects as is usually the case (see Figure 1). For our experiments, we utilized a pre-trained version of YOLO that had been trained on the widely-used COCO dataset. This choice was motivated by the fact that the COCO dataset is one of the largest and most comprehensive datasets available, which allows the YOLO model to achieve high levels of accuracy and robustness across various scenarios.

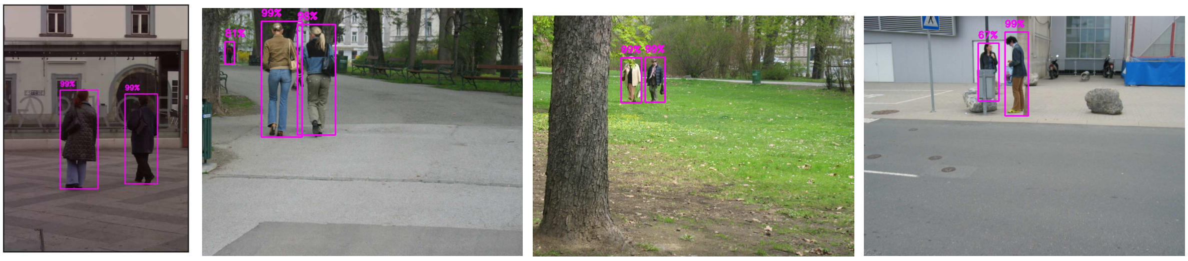

Fig. 1. A sample of the output given by the used vesion of YOLO.

UNet

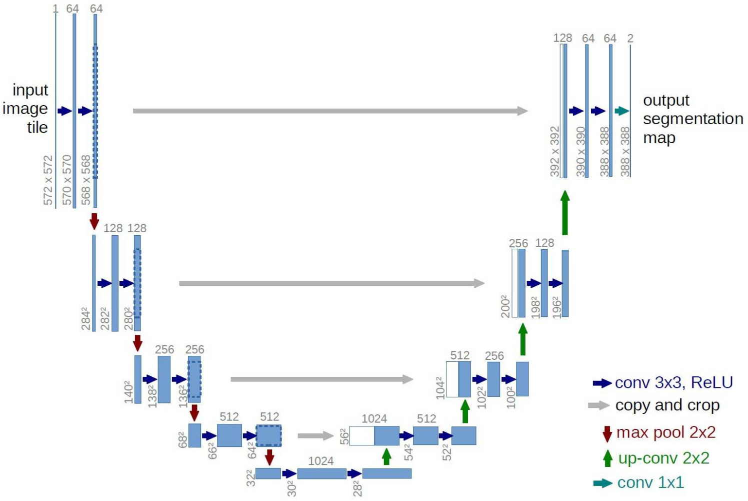

Fig. 2. The UNet architecture [39].

UNet is a very popular model widely used for semantic segmentation tasks in diverse fields such as medical imaging, autonomous driving, and object detection. It has demonstrated exceptional performance in detecting abnormalities, segmenting organs and tissues, and improving decision-making in autonomous vehicles. It is a fully convolutional neural network that takes an image as input and outputs a pixel-wise segmentation map. The UNet architecture consists of a contracting path, which captures the context and reduces the resolution of the input image, and an expanding path, which upsamples the output and fuses it with features from the contracting path to achieve fine-grained segmentation (Figure 2). The UNet model has been shown to perform well on various datasets and is widely used as a baseline model in semantic segmentation tasks.

SqueezeUNet

On the other hand, SqueezeUNet is a neural network architecture that is based on the UNet model, with the added use of Fire units for model compression (Figure 3). The use of Fire blocks enables the network to deal with redundant parameters and improve its overall size by utilizing compact filters for convolutions. This is achieved through the use of two main layers: a "squeeze" layer that utilizes a 1x1 convolutional layer to reduce the number of channels in the input tensor, followed by an "expand" layer that combines 1x1 and 3x3 convolutions to increase the number of channels in the output tensor. This approach effectively reduces the number of parameters, thereby saving on memory and computation costs.

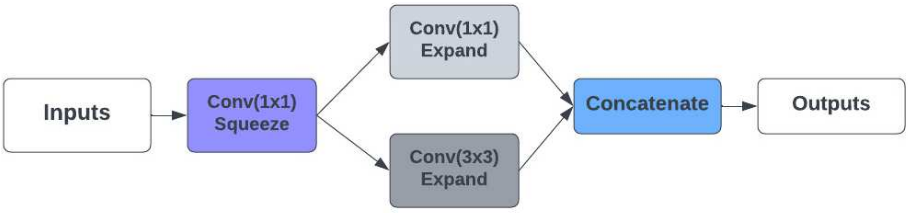

Fig. 3. The fire blocks used in squeezeUNet.

The pipeline

In our proposed pipeline, the input image undergoes initial processing by YOLO, which generates bounding boxes for pedestrians. To enhance the accuracy of the segmentation model in challenging backgrounds and complex scenes, we employ a novel approach wherein we generate separate images for each bounding box instead of directly feeding the bounding box to the segmentation model. To accomplish this, we apply a blur effect to the background surrounding the pedestrian within each bounding box, thereby isolating the pedestrian and its immediate surroundings. By adopting this approach, we avoid any potential distortion of shapes that may arise due to resizing the bounding boxes to fit the input size of the segmentation model. Each of the resulting images is subsequently passed individually to the segmentation model to produce a segmentation mask, which is then placed in its corresponding location in the original image. A schematic representation of this pipeline is provided in Figure 4.

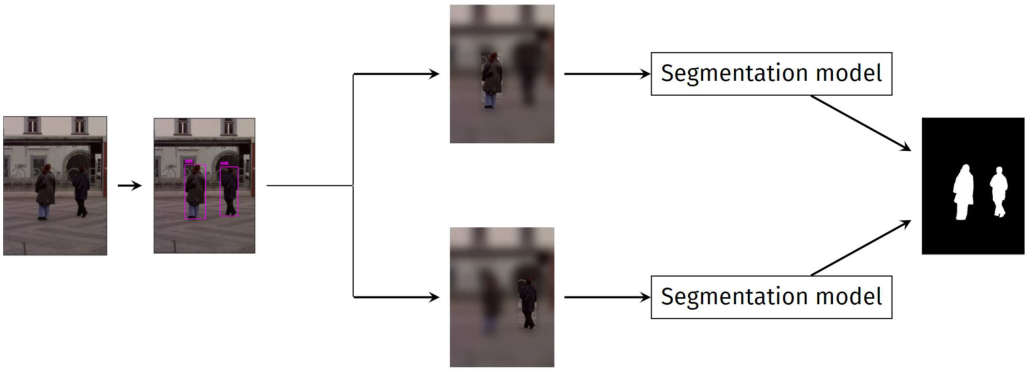

Fig. 4. The pipline of the proposed region proposals-based framework.

OCHuman dataset

The dataset used for training the UNet and SqueezeUNet models is the Occluded Human dataset (OCHuman) [41]. This dataset contains 4731 high-resolution (512 × 512) images with detailed instance masks. The dataset was divided into training and validation sets, with a 70/30 split. The OCHuman dataset was specifically designed for detection-free human segmentation and consists of images of humans in the presence of significant occlusions (a sample of the data is shown in Figure 5). The dataset is particularly challenging due to the presence of occlusions, variations in luminosity, and complex backgrounds. The use of this dataset ensures that the trained models are robust to these challenging conditions and can accurately segment pedestrians in various real-world scenarios. The training was conducted using the same data and parameters on an NVIDIA Tesla K80 GPU and an Intel Xeon 2.3 Ghz CPU.

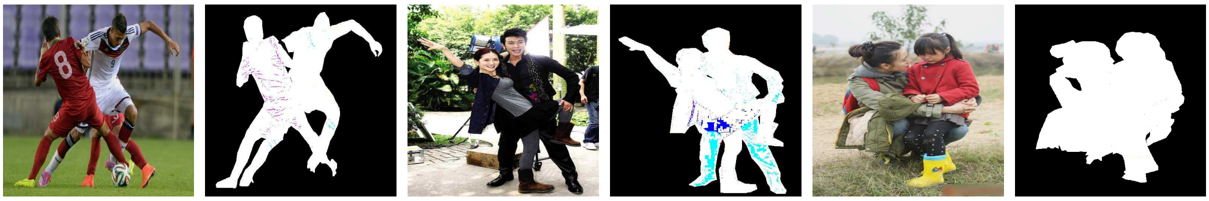

Fig. 5. A sample of OCHuman dataset used to train the models.

3. Experimental results

In this section, we present the experimental results obtained from the proposed framework[1]. We visually compare the segmentation results of UNet, squeezeUNet with and without region proposals. Then, we provide the quantitative evaluation of the models using the Intersection over Union (IoU) and Peak Signal to Noise Ratio (PSNR) which will allow us to compare the accuracy and quality of the segmentation models. The experiments were conducted using an Intel i7 12th generation CPU and an NVIDIA RTX 3050 GPU.

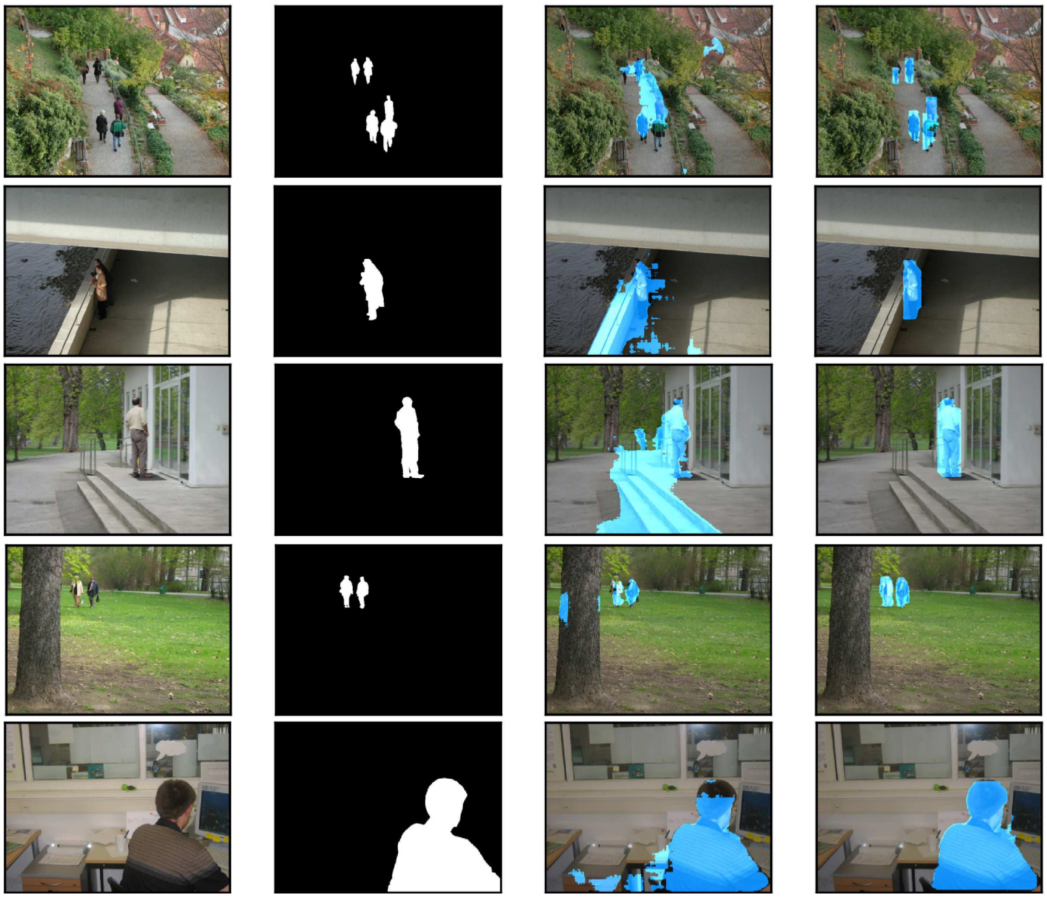

Fig. 6. Comparison of segmentation masks obtained by SqueezeUNet with and without region proposals on a sample of data. The first column shows the input image, the second column shows the ground truth mask, the third column shows the segmentation output obtained by SqueezeUNet without region proposals, and the fourth column shows the segmentation output obtained by SqueezeUNet with region proposals.

In Figure 6, we present a visual comparison of the results obtained by the SqueezeUNet model with and without region proposals. For each row, the first image shows the input image, the second image shows the ground truth mask, the third image shows the mask overlay obtained by the SqueezeUNet model, and the last image shows the mask overlay obtained by the SqueezeUNet model with region proposals. The comparison demonstrates that integrating region proposals significantly improves the segmentation accuracy, reducing the number of false positives in challenging images with occlusions and complex backgrounds. The visual results provide a qualitative confirmation of the effectiveness of our proposed approach.

To evaluate the performance of our proposed approach for pedestrian segmentation quantitatively, we will use two commonly used metrics in image segmentation: Peak Signal-to-Noise Ratio (PSNR) and Intersection over Union (IoU). PSNR is a measure of the reconstruction quality of an image

[1] Source code: https://github.com/MJLahgazi/PedestrianSegRP

compared to a reference image, with higher values indicating better image quality. IoU is a measure of the similarity between the predicted segmentation and the ground truth segmentation, with values ranging from 0 to 1, where 1 represents a perfect match. The equations for PSNR and IoU are defined as follows:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{1}$$

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{255^2}{\mathrm{MSE}} \right), \tag{2}$$

where TP is the number of true positive pixels, FP is the number of false positive pixels, FN is the number of false negative pixels, and MSE is the mean squared error between the predicted and reference images. These metrics will provide quantitative measures of the performance of our proposed approach compared to baseline models.

Table 1 presents a comparison of the performance of SqueezeUNet with and without the proposed region proposal-based framework. The table displays the values of both Peak Signal to Noise Ratio (PSNR) and Intersection over Union (IoU) obtained for a sample of images using both models, the best values are highlighted in bold. Additionally, we report the elapsed time required to obtain the segmentation output for each model. The results clearly demonstrate that the incorporation of region proposals improves the segmentation performance significantly for all the images in the sample. The values of both PSNR and IoU are consistently higher for the SqueezeUNet model with region proposals compared to the baseline model without them. Furthermore, the elapsed time for the proposed framework is comparable to that of the baseline model without region proposals, indicating that our approach does not require significant additional computation time.

Table 1. A sample of the values of PSNR and IoU obtained by SqueezeUNet with and without region proposals.

SqueezeUNet			SqueezeUNet + RP		
PSNR	IoU	Time	PSNR	IoU	Time
12.38	0.74	1.14	**16.99**	**0.90**	1.25
5.64	0.04	0.87	**22.75**	**0.72**	1.27
7.20	0.03	0.85	**25.22**	**0.64**	1.28
9.36	0.17	0.84	**16.73**	**0.69**	1.26
18.87	0.41	0.84	**21.82**	**0.65**	4.35
14.25	0.47	2.06	**15.83**	**0.61**	5.18
10.76	0.19	2.08	**21.02**	**0.72**	3.07
14.14	0.30	1.98	**18.74**	**0.63**	2.40
7.45	0.14	2.11	**18.73**	**0.71**	3.12
23.22	0.28	2.11	**25.86**	**0.57**	3.07
13.14	0.18	2.08	**21.63**	**0.55**	5.10
12.55	0.59	2.16	**17.77**	**0.84**	4.99
12.25	0.66	0.92	**13.67**	**0.76**	3.02
5.79	0.15	0.88	**18.01**	**0.77**	2.09
7.72	0.23	0.83	**17.13**	**0.74**	2.10

In Figure 7, we provide a visual comparison of the segmentation results obtained with the UNet model, following the same layout as in the comparison of squeezeUNet. This comparison illustrates that the integration of region proposals has also had a positive impact on the segmentation performance of the UNet model. The images in the first and second columns represent the input images and the ground truth segmentation masks, while the third and the fourth columns show the segmentation results obtained without and with the use of region proposals, respectively. It is evident from the comparison that the addition of region proposals has led to more accurate and precise segmentation results, with fewer false positives and false negatives. Overall, these results demonstrate the effectiveness of our proposed framework in improving the performance of state-of-the-art segmentation models.

Table 2. A sample of the values of PSNR and IoU obtained by UNet with and without region proposals.

UNet			UNet + RP		
PSNR	IoU	Time	PSNR	IoU	Time
13.52	0.80	1.08	**14.02**	**0.81**	1.46
13.81	0.16	0.95	**26.04**	**0.84**	1.35
14.03	0.09	1.08	**25.07**	**0.63**	2.84
4.35	0.12	2.33	**13.47**	**0.46**	3.32
17.25	0.31	2.80	**22.90**	**0.71**	6.05
11.31	0.24	2.50	**15.27**	**0.60**	5.90
11.11	0.18	2.58	**20.54**	**0.69**	3.34
12.24	0.14	2.31	**18.85**	**0.65**	3.69
9.08	0.18	2.32	**19.13**	**0.69**	3.49
9.84	0.02	2.50	**25.70**	**0.35**	3.44
17.09	0.12	2.45	**21.01**	**0.44**	5.83
10.96	0.47	2.53	**16.74**	**0.81**	5.97
10.64	0.59	2.54	**11.94**	**0.68**	5.44
11.58	0.29	2.35	**17.32**	**0.73**	5.96
5.87	0.16	2.35	**15.23**	**0.64**	5.92

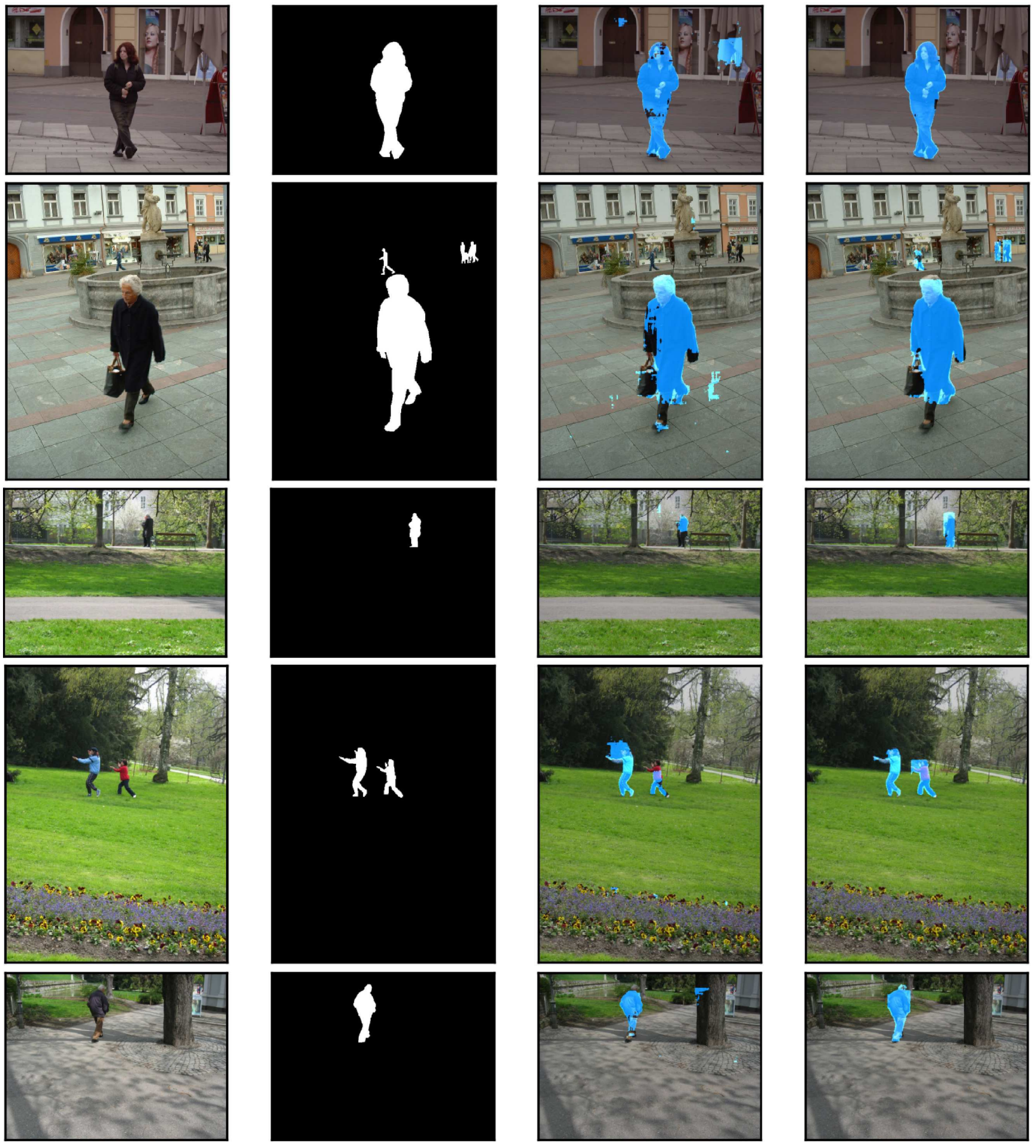

Fig. 7. Comparison of segmentation masks obtained by UNet with and without region proposals on a sample of data. The first column shows the input image, the second column shows the ground truth mask, the third column shows the segmentation output obtained by UNet without region proposals, and the fourth column shows the segmentation output obtained by UNet with region proposals.

Table 2 provides a quantitative analysis of the segmentation performance of UNet, further supporting our earlier conclusions from the experiments with squeezeUNet. The table shows the IoU and PSNR values obtained from the UNet experiments and highlights the best-performing models in bold. These results confirm that integrating region proposals significantly enhances the accuracy of the segmentation model, consistent with our findings from the squeezeUNet experiments.

4. Conclusion

The proposed approach in this paper integrates a region proposal-based framework into the pedestrian segmentation process. By using bounding boxes generated by the YOLO detector as input to the segmentation model, it addresses the challenges of accurately classifying pedestrians in images with challenging backgrounds and occlusions. The experiments conducted on the PASCAL VOC dataset using two different segmentation models, UNet and SqueezeUNet, demonstrate that the incorporation of region proposals significantly improves segmentation accuracy and reduces false positive pixels in the background. The SqueezeUNet model is a compressed network, while the UNet model we used is not as deep. Despite this difference, our proposed approach incorporating region proposals showed significant improvements in segmentation accuracy for both models. The SqueezeUNet and UNet models achieved mIoU values of 0.682 and 0.678, respectively, showing a 12% and 13% improvement over their respective baseline models. The PSNR and IoU values confirmed these findings, and the region proposal framework did not significantly increase the time needed for the segmentation.

[1] Minaee S., Boykov Y. Y., Porikli F., Plaza A. J., Kehtarnavaz N., Terzopoulos D. Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. **44** (7), 3523–3542 (2021).

[2] Hearst M. A., Dumais S. T., Osuna E., Platt J., Scholkopf B. Support vector machines. IEEE Intelligent Systems and their Applications. **13** (4), 18–28 (1998).

[3] Lahgazi M. J., Hakim A., Argoul P. An adaptive wavelet shrinkage based accumulative frame differencing model for motion segmentation. Mathematical Modeling and Computing. **10** (1), 159–170 (2023).

[4] Dalal N., Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). **1**, 886–893 (2005).

[5] Ashok V., Balakumaran T., Gowrishankar C., Vennila I. L. A., Nirmal Kumar A. The Fast Haar Wavelet Transform for Signal & Image Processing. International Journal of Computer Science and Information Security. **7** (2010).

[6] Ren S., He K., Girshick R., Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. **39** (6), 1137–1149 (2015).

[7] Redmon J., Divvala S., Girshick R., Farhadi A. You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 779–788 (2016).

[8] Bochkovskiy A., Wang C.-Y., Liao H.-Y. M. YOLOv4: Optimal Speed and Accuracy of Object Detection. Preprint arXiv:2004.10934 (2020).

[9] Law H., Deng J. CornerNet: Detecting objects as paired keypoints. Proceedings of the European Conference on Computer Vision (ECCV). 734–750 (2018).

[10] Bolya D., Zhou C., Xiao F., Lee Y. J. YOLACT: Real-time instance segmentation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 9156–9165 (2019).

[11] Pavani G., Biswal B., Gandhi T. K. Multistage DPIRef-Net: An effective network for semantic segmentation of arteries and veins from retinal surface. Neuroscience Informatics. **2** (4), 100074 (2022).

[12] Biswal B., Geetha P. P., Prasanna T., Karn P. K. Robust segmentation of exudates from retinal surface using M-CapsNet via EM routing. Biomedical Signal Processing and Control. **68**, 102770 (2021).

[13] Xie H.-X., Lin C.-Y., Zheng H., Lin P.-Y. An UNet-based head shoulder segmentation network. 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW). 1–2 (2018).

[14] Wang P., Bai X. Thermal infrared pedestrian segmentation based on conditional GAN. IEEE Transactions on Image Processing. **28** (12), 6007–6021 (2019).

[15] Baheti B., Innani S., Gajre S., Talbar S. Eff-unet: A novel architecture for semantic segmentation in unstructured environment. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 1473–1481 (2020).

[16] Liu T., Stathaki T. Faster R-CNN for robust pedestrian detection using semantic segmentation network. Frontiers in Neurorobotics. **12**, 64 (2018).

[17] Yuan L., Qiu Z. Mask-RCNN with spatial attention for pedestrian segmentation in cyber-physical systems. Computer Communications. **180**, 109–114 (2021).

[18] Syed A., Morris B. T. CNN, segmentation or semantic embeddings: evaluating scene context for trajectory prediction. International Symposium on Visual Computing. 706–717 (2020).

[19] Gao G., Gao J., Liu Q., Wang Q., Wang Y. CNN-based density estimation and crowd counting: A survey. Preprint arXiv:2003.12783 (2020).

[20] Luo J.-H., Zhang H., Zhou H.-Y., Xie C.-W., Wu J., Lin W. ThiNet: pruning CNN filters for a thinner net. IEEE transactions on pattern analysis and machine intelligence. **41** (10), 2525–2538 (2018).

[21] Reed R. Pruning algorithms-a survey. IEEE Transactions on Neural Networks. **4** (5), 740–747 (1993).

[22] Han S., Pool J., Tran J., Dally W. Learning both weights and connections for efficient neural network. Proceedings of the 28th International Conference on Neural Information Processing Systems. **1**, 1135–1143 (2015).

[23] Li H., Kadav A., Durdanovic I., Samet H., Graf H. P. Pruning filters for efficient convnets. Preprint arXiv:1608.08710 (2017).

[24] He Y., Lin J., Liu Z., Wang H., Li L.-J., Han S. AMC: AutoML for model compression and acceleration on mobile devices. Proceedings of the European conference on Computer Vision (ECCV). 815–832 (2018).

[25] Liu Z., Mu H., Zhang X., Guo Z., Yang X., Cheng K.-T., Sun J. MetaPruning: Meta learning for automatic neural network channel pruning. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 3295–3304 (2019).

[26] He Y., Ding Y., Liu P., Zhu L., Zhang H., Yang Y. Learning filter pruning criteria for deep convolutional neural networks acceleration. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2006–2015 (2020).

[27] Sainath T. N., Kingsbury B., Sindhwani V., Arisoy E., Ramabhadran B. Low-rank matrix factorization for Deep Neural Network training with high-dimensional output targets. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 6655–6659 (2013).

[28] Jaderberg M., Vedaldi A., Zisserman A. Speeding up convolutional neural networks with low rank expansions. Preprint arXiv:1405.3866 (2014).

[29] Denton E. L., Zaremba W., Bruna J., LeCun Y., Fergus R. Exploiting linear structure within convolutional networks for efficient evaluation. Proceedings of the 27th International Conference on Neural Information Processing Systems. **1**, 1269–1277 (2014).

[30] Yin M., Sui Y., Liao S., Yuan B. Towards Efficient Tensor Decomposition-Based DNN Model Compression with Optimization Framework. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 10669–10678 (2021).

[31] Wu B., Wang D., Zhao G., Deng L., Li G. Hybrid tensor decomposition in neural network compression. Neural Networks. **132**, 309–320 (2020).

[32] Bai Z., Li Y., Woźniak M., Zhou M., Li D. DecomVQANet: Decomposing visual question answering deep network via tensor decomposition and regression. Pattern Recognition. **110**, 107538 (2021).

[33] Iandola F. N., Han S., Moskewicz M. W., Ashraf K., Dally W. J., Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. Preprint arXiv:1602.07360 (2016).

[34] Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L. C. MobileNetV2: Inverted residuals and linear bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4510–4520 (2018).

[35] Lee D.-H., Liu J.-L. End-to-end deep learning of lane detection and path prediction for real-time autonomous driving. Signal, Image and Video Processing. **17**, 199–205 (2022).

[36] Chollet F. Xception: Deep learning with depthwise separable convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 1800–1807 (2017).

[37] Wu C. W. ProdSumNet: reducing model parameters in deep neural networks via product-of-sums matrix decompositions. Preprint arXiv:1809.02209 (2018).

[38] Cséfalvay S., Imber J. Self-Compressing Neural Networks. Preprint arXiv:2301.13142 (2023).

[39] Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. **9351**, 234–241 (2015).

[40] Beheshti N., Johnsson L. Squeeze U-Net: A Memory and Energy Efficient Image Segmentation Network. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 1495–1504 (2020).

[41] Zhang S. H., Li R., Dong X., Rosin P., Cai Z., Han X., Yang D., Huang H., Hu S. M. Pose2Seg: Detection Free Human Instance Segmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 889–898 (2019).

Покращення сегментації пішоходів за допомогою семантичної сегментації CNN на основі регіональних пропозицій

Лахгазі М. Дж.[1], Аргул П.[2], Хакім А.[1]

[1] *Факультет наук і техніки, Університет Каді Айяд, Марракеш, Марокко*
[2] *MAST-EMGCU, Університет Гюстава Ейфеля, IFSTTAR, F-77477 Марн-ла-Валле, Франція*

Сегментація пішоходів є критично важливим завданням комп'ютерного зору, але моделям сегментації може бути складно точно класифікувати пішоходів на зображеннях із складним фоном і змінами яскравості, а також оклюзіями. Ця проблема ще більше ускладнюється для стиснутих моделей, які були розроблені для роботи з високими обчислювальними вимогами глибинних нейронних мереж. Щоб вирішити ці проблеми, запропоновано новий підхід, який інтегрує структуру на основі регіональних пропозицій у процес сегментації. Щоб оцінити продуктивність запропонованого фреймворку, проведено експерименти з набором даних PASCAL VOC, який є складним фоном. Використовуємо дві різні моделі сегментації, UNet і SqueezeUNet, щоб оцінити вплив регіональних пропозицій на ефективність сегментації. Проведені експерименти показують, що включення регіональних пропозицій значно покращує точність сегментації та зменшує кількість помилкових пікселів у фоні, що призводить до кращої загальної продуктивності. Зокрема, модель SqueezeUNet забезпечує середнє значення Intersection over Union (mIoU) у розмірі 0.682, що на 12% краще порівняно з базовою моделлю SqueezeUNet без регіональних пропозицій. Подібно модель UNet досягає 0.678, що є покращенням на 13% порівняно з базовою моделлю UNet без регіональних пропозицій.

Ключові слова: *сегментація пішоходів; пропозиції регіону; UNet; виявлення об'єктів; семантична сегментація; згорткові нейронні мережі (CNN).*