**ACCEPTED VERSION**
**2023 IEEE International Symposium on Safety, Security and Rescue Robotics.**
**DOI Pending. Please find the published version on http://ieeexplore.ieee.org.**

# SAR Nets: An Evaluation of Semantic Segmentation Networks with Attention Mechanisms for Search and Rescue Scenes

Andrés Salas-Espinales, Ricardo Vázquez-Martín, Alfonso García-Cerezo, and Anthony Mandow

*Abstract*— **This paper evaluates four semantic segmentation models in Search-and-Rescue (SAR) scenarios obtained from ground vehicles. Two base models are used (U-Net and PSPNet) to compare different approaches to semantic segmentation, such as skip connections between encoder and decoder stages and using a pooling pyramid module. The best base model is modified by including two attention mechanisms to analyze their performance and computational cost. We conduct a quantitative and qualitative evaluation using our SAR dataset defining eleven classes in disaster scenarios. The results demonstrate that the attention mechanisms increase model performance while minimally affecting the computation time.**

*Keywords: Deep Learning; Semantic Segmentation; Attention Mechanism; Disaster Robotics*

## I. Introduction

Natural disasters, human-made accidents, and humanitarian crises are critical events where an effective and fast response is imperative. These scenarios face challenges such as hazardous environments, access limitations, and time constraints. Disaster robotics has emerged in recent decades to increase the capabilities of human rescue teams in tasks such as victim search and assistance [1]–[3] or reduce the risk of first-responders while on duty [4], [5]. Detecting and identifying distinct objects, structures, and potential hazards in complex disaster situations in terrain scenes is a valuable capability for victim search, risk management, and rescue team organization in a Search and Rescue (SAR) mission.

Semantic segmentation involves assigning specific classes to individual pixels within an image. This process has been made efficient through the utilization of convolutional neural networks (CNNs) such as ResNet [6], an architecture featuring five stages and employing residual connections which is widely used as an encoder within semantic segmentation architectures. The architectures of semantic segmentation networks involve reducing image dimensions by extracting features (encoder) and subsequently restoring image spatial size while enhancing semantic-level detection (decoder).

The first semantic segmentation (encoder-decoder) architecture was the Fully Convolutional Network (FCN) [7], which replaces fully connected layers with 1 × 1 convolutional layers, thereby preserving more significant spatial information. U-Net [8] is characterized by the use of skip connections that connect the encoder and decoder to uphold layer-specific information and ensure accurate predictions. SegNet [9], on the other hand, distinguishes from U-Net by storing the location and shape information of detected features in its feature maps. PSPNet [10] introduced a pyramid pooling module that employs parallel clustering layers with diverse filter sizes to capture information at various scales.

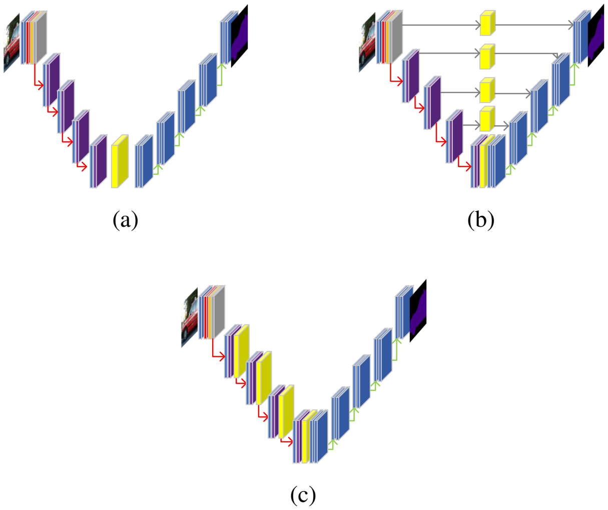
(a)
(b)
(c)

Fig. 1: Model with attention mechanism architectures: (a) between the encoder and the decoder, (b) between multiple skips that connect the encoder and decoder, (c) between each stage of the encoder. The yellow block is the attention mechanism block

To enhance semantic recognition without excessive impact on the computational overhead [11] various researchers have introduced attention mechanisms [12], [13]. This novel method, inspired by cognitive processes, consists of integrating a dedicated attention block between 1) the encoder and the decoder as shown in Figure 1a, 2) multiple skips that connect the encoder and decoder (Figure 1b), or 3) each stage of the encoder (Figure 1c).

For instance, Hu *et al.* [12] introduced the Squeeze and Excitation block (SE), which employs global average pooling (squeeze) to capture contextual information, and a sigmoid activation function (excitation) to model channel-wise dependencies. Another attention mechanism was proposed by Woo *et al.* [13] with their convolutional block attention module (CBAM), which incorporates two distinct attention mechanisms: channel attention, emphasizing crucial information, and spatial attention, focusing on relevant locations within

*This work has been partially funded by the Spanish Ministerio de Ciencia, Innovación y Universidades, Gobierno de España, project PID2021-122944OB-I00.

The authors are with Robotics and Mechatronics Group, Institute of Mechatronics Engineering and Cyber-physical Systems, University of Málaga, Spain. The first author received a grant by Asociación Universitaria Iberoamericana de Postgrado (AUIP), Universidad de Málaga, and Universidad Técnica de Manabí.

an image.

While these models have found application in tasks spanning indoor and outdoor scenes [14], urban environments [15], and synthetic scenarios [16], there remains a need for further research to apply these models to search-and-rescue (SAR) scenes.

Therefore, in this work, we evaluate the performance of two base models (*i.e.*, PSPNet and U-Net) and U-Net with two attention mechanisms (*i.e.*, SE and CBAM) in a SAR dataset, a segmented dataset developed for search-and-rescue purposes. Our main contributions are summarized as follows:

- We add two attention mechanisms to the U-Net model to compare their performance over the U-Net base model.
- We replace the cross-entropy loss function of the original PSPNet and U-Net for two known loss functions (focal loss and dice loss functions).
- We conduct a quantitative and qualitative evaluation of the four models on the SAR dataset.
- We develop a computational analysis for the four models.

These studies show the effectiveness of attention mechanisms to increase the recognition of different classes while minimally affecting computational performance.

The remainder of this article is organized as follows. Section II introduces the SAR dataset used to train the models. Section III describes the architectures of the four models. Section IV specifies the training methodology. Section V shows and discusses the tested results obtained from the four models. Finally, Section VI offers conclusions.

## II. SAR Dataset

To evaluate the behavior of a network in a specific field, a set of images along with their corresponding masks (ground-truth) is needed. Labelled datasets have been created for object detection [17]–[19] and semantic segmentation [20]–[22] in different application domains. In the search-and-rescue (SAR) domain there are a few datasets such as UMA-SAR [23] designed for object detection, VHTA [24] and RescueNet [25] designed for unmanned aerial vehicles (UAV), or DISC [26], a synthetic object detection dataset. However, none of them have segmented masks on terrain scenes.

For this work, we developed a SAR semantic segmentation dataset (see Figure 2), that contains 349 images with their respective hand-labeled segmented annotations split into 70:20:10 ratio (*i.e.*, 70% of images to train, 20% to test, and 10% to validate the models). This dataset has eleven classes in SAR scenarios: first-responder, civilian, vegetation, building, dirt-road, road, sky, civilian car, response-vehicle, debris, and command-post.

## III. Models Description

This section describes the models considered to evaluate their performance on the semantic segmentation SAR dataset: two base models: PSPNet (Figure 3b) and U-Net (Figure 3a), and two models with attention mechanisms (Figure 3c): U-Net-SE and U-Net-CBAM.

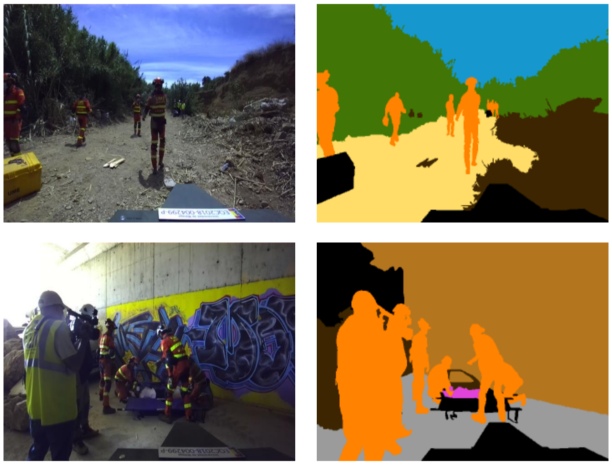

Fig. 2: Two examples from the Semantic Segmentation SAR Dataset: RGB image (left) and ground-truth (right).

The two encoder-decoder base models are designed using as backbone the ResNet-152 network [6] pre-trained on ImageNet [17], where the last average pooling and its fully connected layer are removed. In the case of PSPNet, the pyramid pooling module is added after the previous ResNet-152 stage (*i.e.*, the fifth convolutional layer) as shown in Figure 3b. In the U-Net model, skipped connections are used to concatenate each ResNet-152 stage with its corresponding decoder stage, as shown in Figure 3a. Both models use the same decoder architecture (*i.e.*, each decoder stage is up-sampled by two to restore the original pixels dimension for pixel-wise classification).

With respect to U-Net with attention mechanisms illustrated on the Figure 3c, the attention mechanism module can be SE and CBAM. The module is added after the skip connections of each encoder-decoder stage (*i.e.*, between each encoder stage). This is done to preserve the original feature map of the encoder stage that has to be concatenated with its corresponding decoder stage.

## IV. Training methodology

### A. Implementation details

The models were implemented on a DGX station with one NVIDIA Tesla® V100 32GB GPU using Pytorch 1.3 toolbox. We used the stochastic gradient descent (SGD) with a learning rate of 0.04, momentum of 0.9, weight decay of 0.0005, and a batch size of five. We resized the images to 480x640 pixels and for data augmentation, all training images were flipped, cropped, and cropped out, in addition, brightness was applied to all the trained images. Each model was trained using the early stopping technique [27] with a patience of 40.

The original PSPNet and U-Net use the well know cross-entropy loss function. Nevertheless, in this work we adapt these networks by using two loss functions added together: the dice loss [28] and the focal loss [29] functions, where each of them has a weight of 0.5. Both loss functions

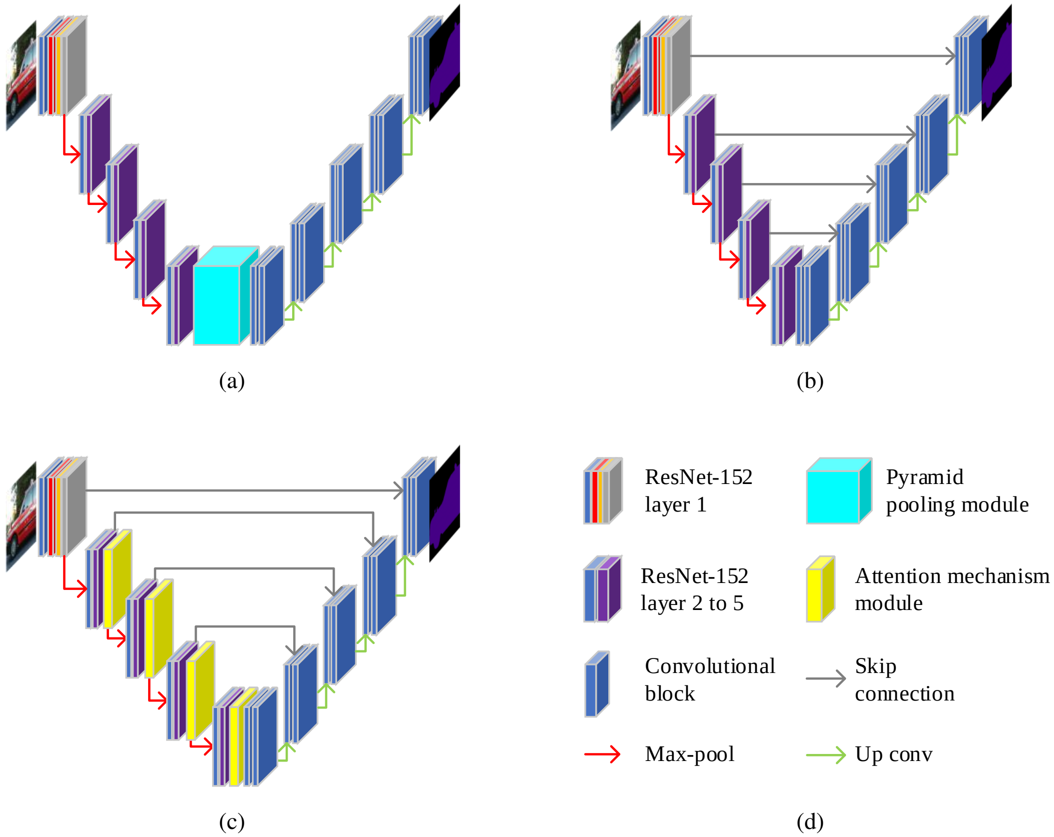

Fig. 3: Model architectures: (a) PSPNet, (b) U-Net, (c) U-Net with SE or CBAM attention mechanisms, and (d) Nomenclature

were designed for highly unbalanced datasets (as our SAR dataset).

### B. Evaluation metric

To evaluate the performance of the networks, the mean accuracy ($mAcc$) and the intersection-over-union ($mIoU$) are used as precision scores. To evaluate the computational performance we measure the floating point operations per second (FLOPS), the number of parameters, inference time in milliseconds (ms), and inference speed in frames per second (fps).

## V. Results and Discussion

This section discusses quantitative and qualitative results obtained from the training process. It also presents a computational and time complexity analysis that illustrates the number of FLOPS, parameters, inference time, and inference speed consumed by the models.

### A. Quantitative results

Table I shows the quantitative results of the two base models (*i.e.*, PSPNet, and U-Net), and the U-Net model with the two attention mechanisms (*i.e.*, SE, and CBAM). First, the results obtained in the base models are analyzed. Next, the scores between the U-Net base model and U-Net with both attention mechanisms are compared. Then, the values obtained from U-Net-SE and U-Net-CBAM are analyzed. Finally, a general overview is given.

As for the base model comparisons, Table I shows that the $IoU$ of U-Net is better in all the classes where the more significant scores are obtained from first-responder class (approximately 12% greater), and both the response-vehicle and command-post classes (around 5% greater). As a result, the $mIoU$ score of U-Net is 3.5% greater than PSPNet. In addition, the results show that the $Acc$ of U-Net is better in almost all the classes with the exception of the dirt-road class which is 1% less than the PSPNet model. The more significant classes predicted by U-Net are first-responder (around 7%), road (around 5%), and response-vehicle (around 4%). Due to this, the U-Net $mAcc$ score is approximately 3% better than PSPNet. These scores indicate that the most suitable model for this kind of dataset (*i.e.*, small and unbalanced dataset) is U-Net with its skip connections between encoder-decoder layers, compared with the pyramid pooling module of PSPNet.

Regarding the scores obtained from the U-Net base model and U-Net with an attention mechanism, they show that, as expected, U-Net with an attention mechanism increases the prediction percentage in almost all the classes with the exception of the vegetation class (where in U-Net-SE is around 0.2 % lower and in U-Net-CBAM is around 0.8 % lower). The classes that considerably increase their predictions in U-Net-SE are civilian cars (approximately 5%) and debris (approximately 3%), and in U-Net-CBAM are road (approximately 5%) and dirt-road (approximately 2%).

TABLE I: Quantitative results(%) on the SAR test set. Values in red indicate the best results.

| Classes | Base models | | | | U-Net with attention mechanisms | | | |
|---|---|---|---|---|---|---|---|---|
| | PSPNet | | U-Net | | U-Net-SE | | U-Net-CBAM | |
| | $Acc$ | $IoU$ | $Acc$ | $IoU$ | $Acc$ | $IoU$ | $Acc$ | $IoU$ |
| Background | 60.90 | 51.37 | 67.34 | 57.32 | 68.39 | 57.16 | 70.14 | 57.73 |
| First-responder | 85.19 | 67.24 | 91.99 | 79.41 | 91.70 | 79.99 | 91.41 | 80.14 |
| Civilian | 00.00 | 00.00 | 00.00 | 00.00 | 00.00 | 00.00 | 00.00 | 00.00 |
| Vegetation | 88.04 | 74.59 | 90.14 | 77.62 | 91.40 | 77.47 | 90.89 | 76.99 |
| Road | 57.26 | 49.77 | 62.59 | 51.97 | 60.99 | 52.19 | 64.28 | 56.43 |
| Dirt-road | 87.36 | 75.20 | 86.90 | 76.18 | 86.51 | 76.57 | 87.32 | 77.78 |
| Building | 93.21 | 88.33 | 95.58 | 90.69 | 94.86 | 91.25 | 95.64 | 91.43 |
| Sky | 98.84 | 94.84 | 99.12 | 95.81 | 99.14 | 96.01 | 99.18 | 95.93 |
| Civilian car | 93.25 | 81.38 | 96.42 | 83.98 | 95.33 | 88.85 | 95.49 | 84.81 |
| Response-vehicle | 65.18 | 59.91 | 69.16 | 64.57 | 69.11 | 64.21 | 68.86 | 65.01 |
| Debris | 82.29 | 71.67 | 82.72 | 73.76 | 85.16 | 76.27 | 81.59 | 74.10 |
| Command-post | 96.03 | 86.11 | 97.30 | 91.13 | 96.78 | 91.23 | 98.07 | 90.93 |
| | $mAcc$ | $mIoU$ | $mAcc$ | $mIoU$ | $mAcc$ | $mIoU$ | $mAcc$ | $mIoU$ |
| | 75.63 | 66.70 | 78.27 | 70.20 | 78.28 | 70.93 | 78.57 | 70.94 |

Because of those predictions, the $mIoU$ scores of U-Net with attention mechanisms increase by 0.7% approximately with respect to the U-Net base model. This increment is related to the main objective of attention mechanisms which is allowing the network to focus on important features and learn more discriminate representations [11].

A comparison between U-Net-SE and U-Net-CBAM shows that in reference to $IoU$ values, on the one hand, U-Net-SE outperforms in the civilian car class with around 4%, debris with around 2%, and in classes such as vegetation, sky, and command-post with a slight increase of around 0.5%; on the other hand, U-Net-CBAM outperforms in five classes: road with around 4%, dirt-road and responder-vehicle with around 1%, and in building and first-responder in around 0.2%. These balance $IoU$ predictions result in similar $mIoU$. The same behavior occurs with $Acc$, resulting in a slight $mAcc$ difference of 0.29% favorable to U-Net-CBAM.

In contrast to the classes mentioned above, the civilian class shows no recognition neither in $Acc$ and $IoU$, which can be explained by two reasons: the similarity that this class has with the first-responder class and the unbalance of the dataset.

All in all, the use of attention mechanisms improves the accuracy for all classes, with the exception of vegetation, which offers a very similar metric. Furthermore, both attention mechanisms offer equivalent $mAcc$ and $mIoU$, even if U-Net-CBAM outstands for classes related to autonomous navigation such as road, dirt-road, and response vehicles. Conversely, U-Net-SE outstands for situational awareness purposes in SAR scenarios with classes such as civilian-car, command-post, or debris.

### B. *Qualitative visualization of experiments*

We carry out a qualitative analysis of the evaluated methods by comparing the ground-truth, the predicted images from PSPNet, U-Net, U-Net-SE, and U-Net-CBAM for five images from SAR dataset as shown in Figure 4.

All the evaluated models detect almost all the classes with the exception of the civilian class. As there was mentioned before it can be due to the similarity that this class has with the first-responder class as shown in the figure of the first row (*i.e.*, the ground-truth defines the person as a civilian class but the models predicted it as a first-responder).

A comparison between PSPNet and U-Net predictions indicates the effectiveness of U-Net to define better boundaries for the predicted classes. Additionally, U-Net with both attention mechanisms defines boundaries slightly better than U-Net alone.

Moreover, the examples also present classes that are detected better in U-Net-SE than U-Net-CBAM such as the third and fourth rows where they show a better detection of the response-vehicle and the debris classes. In contrast, the fourth row shows images where the dirt-road class is better defined by U-Net-CBAM rather than U-Net-SE. Thus, these visual comparisons reinforce the scores obtained in Table I.

The final example (*i.e.*, the fifth row) visually indicates the importance of attention mechanisms, it shows the correct prediction of air cables as background which are not even segmented in the ground-truth.

### C. *Computational performance analysis*

Table II presents the results related to the computational performance where PSPNet compared with U-Net has better results in the number of FLOPS (around half of U-Net), and the number of parameters (around 15M less than U-Net); however, U-Net shows better performance in the inference speed, inference time, $mAcc$, and $mIoU$. Therefore, It can be said that in this specific task (SAR dataset), skips connections worsen the computational cost of a model but outperforms the time complexity and the prediction scores compared with the use of a pyramid pooling module.

As for the use of attention mechanisms, both SE and CBAM slightly increase the computing consumption and time complexity but enhance the prediction scores. Therefore, for this specific task, it is important to consider the addition of attention mechanisms to base models. It is true

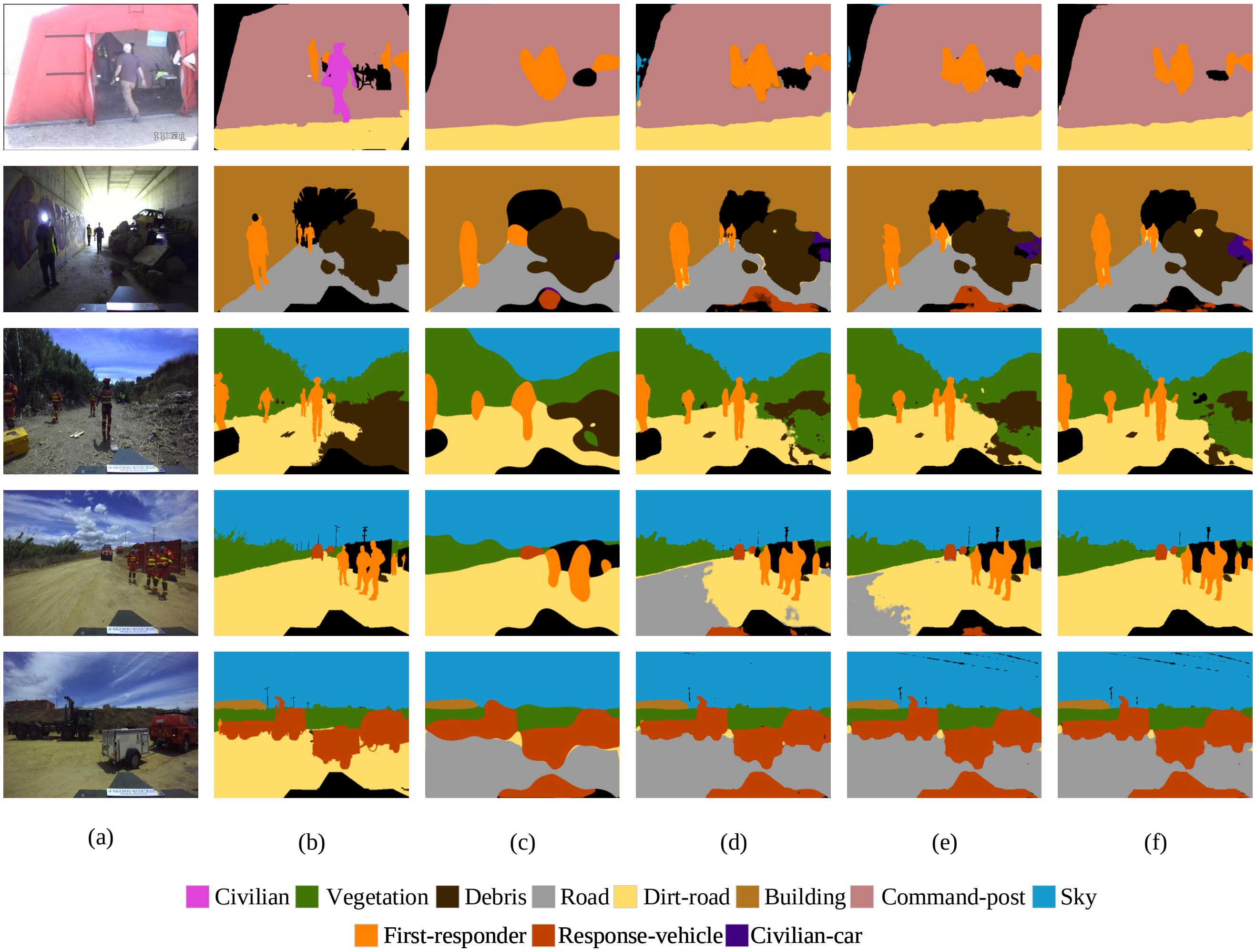

Fig. 4: Predictions on SAR dataset: (a) RGB images, (b) Ground-truth images, (c) PSPNet, (d) U-Net, (e) U-Net-SE, and (f) U-Net-CBAM.

TABLE II: Computing and time complexity analysis on SAR test set. The inference speed was measured by running a test on a single GPU

| Network | FLOPS (G) | Params (M) | Inference speed (fps) | Inference time (ms) | $mAcc$ | $mIoU$ |
|---|---|---|---|---|---|---|
| PSPNet | 108.74 | 91.72 | 32.27 | 30.99 | 75.63 | 66.70 |
| U-Net | 219.46 | 106.51 | 34.49 | 29.00 | 78.27 | 70.20 |
| U-Net-SE | 219.47 | 107.20 | 31.20 | 32.06 | 78.28 | 70.93 |
| U-Net-CBAM | 219.47 | 107.20 | 31.40 | 31.84 | 78.57 | 70.94 |

that the values of the computational and time model cost are penalized; however, it significantly increases the prediction scores of specific classes for SAR scenarios such as debris, civilian car, and command-post.

## VI. Conclusions

This paper presents an evaluation of four semantic segmentation models in Search-and-Rescue (SAR) scenarios. Two different base models corresponding to different semantic segmentation approaches have been used: U-Net, based on skip connections between encoder and decoder, and PSPNet, which rely on a pyramid pooling module. The best of the two base models is modified by integrating two different attention mechanisms, the convolutional block attention module (CBAM) and the Squeeze and Excitation block (SE). To evaluate their performance the mean accuracy ($mAcc$) and the intersection-over-union ($mIoU$) are used as precision scores, and to evaluate the computational performance we measure the floating point operations per second (FLOPS), the number of parameters, inference time in milliseconds (ms), and inference speed in frames per second (fps). The four models have been trained on our manually labeled SAR dataset with eleven classes. The quantitative and qualitative results show a better performance of the U-Net model, detecting classes with more accurate boundary precision, and the effectiveness of attention mechanisms to increase the model performance while minimally affecting the computational cost.

As future work, we are interested in augmenting and balancing the SAR dataset, and including different visual modalities such as thermal infrared or depth information.

## Acknowledgments

The authors want to thank the collaboration of the Chair for Safety, Emergencies, and Disasters of Universidad de Málaga. In addition, we would like to thank all the members of the Robotics and Mechatronics Lab, for obtaining ground images during SAR exercises. This work has been partially funded by the Ministerio de Ciencia, Innovación y Universidades, Gobierno de España, project PID2021-122944OB-I00. The first author received a grant from Asociación Universitaria Iberoamericana de Postgrado (AUIP), Universidad de Málaga, and Universidad Técnica de Manabí.

## References

[1] F. Pastor, F. J. Ruiz-Ruiz, J. M. Gómez-de Gabriel, and A. J. García-Cerezo, "Autonomous wristband placement in a moving hand for victims in search and rescue scenarios with a mobile manipulator," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 871–11 878, 2022.

[2] M. Toscano-Moreno, J. Bravo-Arrabal, M. Sánchez-Montero, J. Serón-Barba, R. V. Martín, J. J. F. Lozano, A. Mandow, and A. García-Cerezo, "Integrating ROS and android for rescuers in a cloud robotics architecture: Application to a casualty evacuation exercise," in *IEEE International Symposium on Safety, Security, and Rescue Robotics, SSRR 2022, Sevilla, Spain, November 8-10, 2022*. IEEE, 2022, pp. 270–276. [Online]. Available: https://doi.org/10.1109/SSRR56537.2022.10018629

[3] R. Edlinger, C. Föls, and A. Nüchter, "An innovative pick-up and transport robot system for casualty evacuation," in *IEEE International Symposium on Safety, Security, and Rescue Robotics, SSRR 2022, Sevilla, Spain, November 8-10, 2022*. IEEE, 2022, pp. 67–73. [Online]. Available: https://doi.org/10.1109/SSRR56537.2022.10018818

[4] A. Seino, N. Seto, L. Canete, and T. Takahashi, "Long-reach compact robotic arm with lmpa joints for monitoring of reactor interior," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 6384–6389.

[5] D. De Schepper, I. Dekker, M. Simons, L. Brabants, W. Schroeyers, and E. Demeester, "Towards a semi-autonomous robot platform for the characterisation of radiological environments," in *2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, 2022, pp. 230–237.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[7] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, 2015, pp. 234–241.

[9] 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.

[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6230–6239.

[11] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Computer Vision – ECCV 2018*, 2018, pp. 3 – 19.

[14] M. Fan, J. Lu, X. Kong, W. Sun, W. Sun, and Y. Sun, "Cbam-dce: A non-reference image correction algorithm for uneven illumination," in *Proceedings of the International Conference on Artificial Intelligence and Pattern Recognition*, 2023, p. 803–809.

[15] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 593–602.

[16] F. Li, Z. Jiang, S. Zhou, Y. Deng, and Y. Bi, "Spilled load detection based on lightweight yolov4 trained with easily accessible synthetic dataset," *Computers and Electrical Engineering*, vol. 100, p. 107944, 2022.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, 2014, pp. "740–755.

[19] Z. Dong, K. Xu, Y. Yang, H. Bao, W. Xu, and R. W. Lau, "Location-aware single image reflection removal," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5017–5026.

[20] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

[21] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 891–898.

[22] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5122–5130.

[23] A. Banuls, A. Mandow, R. Vazquez-Martin, J. Morales, and A. Garcia-Cerezo, "Object detection from thermal infrared and visible light cameras in search and rescue scenes," in *IEEE International Symposium on Safety, Security, and Rescue Robotics*, 2020, pp. 380 – 386.

[24] X. Zhang, Y. Feng, S. Zhang, N. Wang, S. Mei, and M. He, "Semi-supervised person detection in aerial images with instance segmentation and maximum mean discrepancy distance," *Remote Sensing*, vol. 15, no. 11, 2023.

[25] T. Chowdhury, R. Murphy, and M. Rahnemoonfar, "Rescuenet: A high resolution uav semantic segmentation benchmark dataset for natural disaster damage assessment," 2022.

[26] H. G. Jeon, S. Im, B. U. Lee, F. Rameau, D. G. Choi, J. Oh, I. S. Kweon, and M. Hebert, "A large-scale virtual dataset and egocentric localization for disaster responses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[27] T. Li, Z. Zhuang, H. Liang, L. Peng, H. Wang, and J. Sun, "Self-validation: Early stopping for single-instance deep generative priors," in *32nd British Machine Vision Conference*, 2021, p. 108. [Online]. Available: https://www.bmvc2021-virtualconference.com/assets/papers/1633.pdf

[28] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2017, pp. 240–248.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.