36th European Modeling & Simulation Symposium, 037
21th International Multidisciplinary Modeling & Simulation Multiconference

2724-0029 © 2024 The Authors.
doi: 10.46354/i3m.2024.emss.037

Enhancing Defense Threat Assessment through Segmentation and Labelling of Synthetic Data

Agostino Bruzzone[1,*], Umberto Battista[2], Carolina Badano[2] & Federico Taddei Santoni[2]

[1]Simulation Team & Genoa University
[2]Stam S.r.l., Via Pareto 8 AR, 16129, Genoa, Italy

*Corresponding author. Email address: agostino.bruzone@simulationteam.com

Abstract

Threat assessment requires a nuanced understanding of potential risks and vulnerabilities across various scenarios. It is imperative to identify, analyse, and mitigate potential risks effectively due to the uncertainty that appears largely in the landscape of threats. This article explores the complex task of threat assessment, particularly in the military domain, where the consequences of threats could be catastrophic. Military threat assessment is traditionally based on expert judgment, historical data analysis, and simulation techniques. However, in data-scarce environments like the military, innovative approaches leveraging artificial intelligence, natural language processing, and anomaly detection have emerged as invaluable tools. This article explores the integration of segmentation, labelling, and contextual understanding on real and synthetic images in order to perform a thorough threat assessment, fruitful for decision-making process. It presents advanced techniques such as R-CNN, Mask R-CNN, HCP, CLIP, and Mask CLIP+. Through a comprehensive review and analysis, the article highlights the significance of these techniques in enhancing the accuracy, efficiency, and depth of threat analysis in military operations.

Keywords: Threat Assessment, Segmentation, Labelling, Artificial Intelligence, Defence, Synthetic data, Simulation

1. Introduction

A threat is a source of potential harm that could happen due to vulnerabilities present in a scenario. One of the key elements affecting threats is that are affected by VUCA (Volatility, Uncertainty, Complexity and Ambiguity). In facts, usually, there is no certainty about what will happen, when it will happen, how bad the consequences will be, how long it will last and of the probability of it happening; indeed, usually threats are characterized by using the VUCA factors to increase their capability to overpass defences and succeed in addressing vulnerabilities. Therefore, for defense of a System one of the crucial aspect is to identify and classify all possible threats and to develop approaches to detect and assess them as soon as possible to increase probability to deflect or neutralize, as well as to mitigate consequences. In this process the crucial aspect is to detect also vulnerabilities and reduce them in order to increase resilience of the systems, that therefore relies on the capability to react dynamically to threats and self reorganize to face them. Within this context the threat assessment techniques are used, particularly in the military field where the occurrence of a threat has higher probability with potential catastrophic effects. From this point of view the adoption of Strategic Engineering approach combining AI, M&S and Data Analytics (Bruzzone, 2018); indieed this integration is a crucial aspect to fill up the gaps related to gaps such lack of data or inconsistencies. Indeed, from this point of view it turns crucial to be able to address the Threat Assessment that is the process of elaborating data and information to evaluate the

threats and use the related results in decision making for evaluating most appropriate priorities and reactions. Hence, threat assessment is defined as the practice of determining the credibility and severity of potential threats, as well as the likelihood that the threat will become a reality, by extracting information from available data, boundary conditions and all environmental parameters as well as analysing behaviours (Waltz & Llinas, 1990; Bruzzone et al., 2014).

Identifying potential threats is a complex task influenced by various factors. Institutions, particularly in the military field, play a significant role in shaping threat assessments by providing guidelines for organised human interaction across different domains like families, governments, businesses, and religions. Cognitive influence adds another layer of complexity, making prediction and assessment challenging. Risk, being dynamic and immeasurable, requires continuous evaluation. Choosing the alternative with the lowest expected risk is often rational when probabilities are comparable. However, challenges arise in assessing probabilities and comparing consequences, particularly in analyses focusing on conceptual and psychological outcomes.

Conducting risk assessment in situations where there is a lack of data is challenging; indeed, in such cases, it is necessary to rely on alternative methods and strategies to gather relevant information and make informed assessments. Traditional methods such as relying on expert judgement, and analysing historical data form the basis of military threat assessment. Simulation and modelling techniques contribute by allowing the testing of different scenarios, providing a nuanced understanding of potential consequences and assessing the resilience of military systems (Amico, et al., 2000; Bruzzone & Massei 2017). In data-scarce situations, generative learning techniques such as artificial intelligence, natural language processing, and anomaly detection greatly improve military risk assessment. These techniques make it easier to create artificial data for training, detect irregular patterns that indicate potential threats, create a variety of threat scenarios for training and readiness, improve current data through variation, apply knowledge from related tasks to improve threat recognition, summarize complex information for effective threat communication, work with generative models and human analysts to create scenarios, update threat models continuously based on new data and information.

In this paper the main focus is on visual data and the capability to develop solutions able to extract information and knowledge by the related channels such as EO/IR to understand features, behaviours and criticalities, identify symptoms and detect threats (Sizintsev, et al., 2019; Blasch et al., 2021; Vakil et al., 2021). Indeed, the paper address the issue of how to visualise potential threat landscapes through immersive simulations.

2. State of the art on Visual Threat Assessment

Performing threat assessments relies on identifying potential threats, which is facilitated by segmentation and labelling. Segmentation involves breaking down the system into distinct parts, while labelling assigns names and classifications to each component. This structured approach aids both human analysts and machine learning algorithms in understanding and responding to security risks effectively.

2.1. Segmentation

Image segmentation involves dividing the image into distinct and semantically meaningful parts, such as objects, backgrounds or specific features. Several algorithms and techniques are used to perform segmentation. In conditions of data scarcity, segmentation could be utilized in synthetic scenes to identify various aspects of military environments, including terrain, infrastructure, and objects of interest, with fidelity. This segmentation process entails breaking down synthesized imagery into distinct and semantically meaningful parts, such as vehicles, buildings, vegetation, and terrain features.

2.1.1. *R-CNN*

Region-based Convolutional Neural Network (R-CNN) is a computer vision algorithm used for object detection. It works by first proposing regions in an image that might contain objects using selective search or similar methods. Then, it extracts features from each proposed region using a Convolutional Neural Network (CNN). These features are fed to a set of support vector machines to classify and refine the proposed regions. Finally, non-maximum suppression is applied to generate the final bounding boxes for detected objects.

Mask R-CNN, signifying a conceptual extension of Faster R-CNN, enriches its capabilities by introducing a third branch dedicated to predicting segmentation masks for each candidate object. Essentially, Faster R-CNN initially provides class labels and bounding-box offsets for each candidate object, and Mask R-CNN seamlessly incorporates an additional branch to furnish object masks. This intuitive extension facilitates a finer spatial layout extraction, a pivotal requirement for accurate segmentation. The foundational structure of Mask R-CNN derives from the two-stage framework of Faster R-CNN. The initial stage involves the Region Proposal Network (RPN), proposing bounding boxes for candidate objects. Subsequently, akin to Fast R-CNN, the second stage extracts features using Region of Interest (RoI) Pooling from each candidate box, facilitating classification and bounding-box regression. A distinctive feature of Mask R-CNN's second stage is its concurrent output of binary masks for each RoI, deviating from recent

systems where classification depends solely on mask predictions. The use of ResNet as the backbone is a strategic move to address concerns related to network depth, calculation, and parameter quantity. Additionally, Group Normalization (GN) is introduced to enhance detection accuracy, and cascade training with Intersection over Union (IoU) thresholds is employed to further refine detector performance. This proposed algorithm strategically tackles the limitations of Mask R-CNN, ensuring accurate bounding box and mask information for subsequent classification and regression tasks. The combination of Mask R-CNN's innate capabilities with these enhancements underscores the adaptability and effectiveness of this comprehensive approach.

2.2. Labelling

The labelling phase occurs after the segmentation process, focusing on the segmented images and various objects within them. This sequential approach allows for a more targeted and refined analysis of individual objects. Thus, labelling images offers a comprehensive understanding of their content, a crucial factor for accurate threat assessment. It enhances the interpretation of image content by associating meaningful categories, aids in precise identification and classification of individual objects, adds context to segmented elements for better scenario analysis, serves as valuable training data for machine learning models, generates easily understandable results for collaboration with experts, supports informed decision-making across various domains, extracts actionable information tailored to specific use cases, and contributes to visually informative representations for improved analysis. In the context of the military field characterized by scarcity of data, labelling within synthetic data plays a critical role in identifying and categorizing potential threats, enabling comprehensive threat assessment and readiness training. By accurately labelling objects within synthetic imagery, analysts could effectively evaluate the nature and severity of simulated threats, enhancing preparedness and response capabilities in military operations.

In the realm of labelling within computer vision, Convolutional Neural Networks play a fundamental role in extracting relevant features and making predictions based on visual data. A CNN is a specialized type of neural network designed to process grid-like data, such as images. Its architecture includes convolutional layers that apply filters to input data, enabling the network to automatically learn hierarchical representations of features.

2.2.1. *HCP*

An innovative solution for multi-label classification is the Hypotheses-CNN-Pooling (HCP) structure, a flexible deep CNN architecture. HCP efficiently processes an arbitrary number of object segment hypotheses, potentially generated by advanced objectiveness detection techniques. Each hypothesis is seamlessly connected to a shared CNN, and a novel pooling layer is introduced for aggregating single-label CNN predictions into multi-label results. HCP does not demand ground-truth bounding box information during training on multi-label image datasets. Additionally, to address potentially noisy hypotheses, HCP employs a cross-hypothesis max-pooling operation, effectively suppressing noise and discarding redundant hypotheses. Moreover, the shared CNN within HCP is flexible, allowing pre-training with large-scale single-label image datasets like ImageNet. Fine-tuning on the target multi-label dataset is facilitated. Lastly, HCP's outputs, processed through the softmax layer, result in normalized probability distributions over labels. The Hypotheses-CNN-Pooling deep network is designed to handle multi-label image classification, addressing challenges such as the absence of ground-truth bounding box information and the need for robustness to noisy or redundant hypotheses.

2.2.2. *CLIP*

Large-scale visual-language pre-training models, exemplified by CLIP, excel at capturing rich and expressive features in both visual and language domains. The utilization of raw CLIP features for zero-shot image classification proves to be a robust and competitive strategy, demonstrating performance comparable to fully-supervised counterparts. CLIP learns from images of complex scenes and their accompanying natural language descriptions. This unique learning paradigm encourages the embedding of local image semantics in its features. Moreover, it empowers CLIP to learn concepts in an open vocabulary, accommodating a wide range of objects and capturing rich contextual information. Notably, CLIP's ability to grasp the co-occurrence and relations of certain objects, along with spatial priors, contributes to its versatility in diverse tasks. Its architecture consists of an image encoder and a text encoder, both jointly trained to map input images and text into a unified representation space. The training objective employs contrastive learning, treating ground-truth image-text pairs as positives and creating negatives from mismatched image-text combinations. CLIP offers two alternative implementations: a Transformer and a ResNet with a global attention pooling layer.

MaskCLIP simplifies dense patch-level feature extraction from CLIP's image encoder while preserving visual-language associations. It leverages classification weights directly from CLIP's text encoder embeddings, utilizing 1×1 convolutions without explicit mapping. Compatibility extends to all CLIP variants, including ResNets and ViTs. Key smoothing and prompt denoising refine MaskCLIP's performance without training. MaskCLIP+ integrates into training processes, providing high-quality pseudo labels for advanced segmentation models like PSPNet and

DeepLab. It adapts to various semantic segmentation scenarios, including open-vocabulary and fine-grained class segmentation, retaining CLIP's robustness against distribution shifts and corruptions. MaskCLIP+ extends to transductive zero-shot segmentation, generating pseudo labels for unseen classes. Unlike object detection approaches, MaskCLIP+ relies on pseudo labels, ensuring consistent performance across seen classes.

3. Comparative Analysis

RCNN is primarily designed for object detection tasks. It operates by proposing regions of interest within an image and then classifying these regions into different object categories. While effective for detection, RCNN does not provide segmentation or labelling capabilities directly. Instead, mask RCNN extends RCNN by adding a segmentation component to it. Along with object detection, it is also possible to generate pixel-level segmentation masks for each detected object. This means it not only identifies objects but also precisely delineates their boundaries.

HCP is specialized in human-centric labelling tasks. It is designed to recognize and label various parts of the human body in images. Unlike RCNN and Mask RCNN, which are more general-purpose, HCP is tailored specifically for human-related labelling tasks.

CLIP is a model capable of understanding and labelling images based on natural language descriptions. It is able to associate textual descriptions with visual content, enabling it to perform tasks like image labelling. However, it does not provide segmentation capabilities like Mask RCNN. To perform segmentation it could be used also MaskCLIP, an extension of CLIP. It is able to label images and provide segmentation masks for objects within those images. This allows for more detailed understanding and analysis of visual content compared to CLIP alone. MaskCLIP+ is an improvement over MaskCLIP, focusing on producing higher quality segmentation masks. It enhances the segmentation accuracy and quality, providing more precise delineation of object boundaries compared to its predecessor.

A benchmarking analysis is carried out to rate the above methods between 1 and 5, considering the following evaluation criteria: accuracy, model precision; robustness, stability across different conditions; ease of use is the measure about how easily it is possible to install, integrate and use the model in workflows; flexibility corresponds to the adaptability to different tasks, while the feasibility correspond to how much the model is practical to use in terms of computational resources.

4. Results and Discussion

Mask R-CNN provides a comprehensive solution for threat assessment tasks due to its ability to perform both object detection and pixel-level segmentation. This makes it well-suited for scenarios requiring detailed segmentation of threat objects, enabling precise delineation of object boundaries for analysing potential threats within images.

Mask CLIP and its enhanced version, Mask CLIP+, offer a unique combination of labelling, segmentation, and contextual understanding based on natural language descriptions. These models provide a holistic approach to threat assessment by integrating segmentation capabilities with contextual understanding, potentially improving the accuracy and efficiency of threat identification.

Considering the benchmarking analysis shown in Table 1, which evaluates accuracy, robustness, ease of use, flexibility, and feasibility, along with the requirement for both segmentation and labelling in threat assessment tasks, the most promising method is Mask CLIP+. It combines state-of-the-art segmentation capabilities with contextual understanding, offering a comprehensive solution for threat assessment tasks.

Figure 1. Benchmarking analysis

	Accuracy	Robustness	Ease of use	Flexibility	Feasibility	TOT
RCNN	4	3	3	3	3	16
Mask RCNN	5	4	4	4	4	21
HCP	4	4	4	2	4	18
CLIP	3	3	3	4	3	16
Mask CLIP	4	4	4	5	4	21
Mask CLIP+	5	5	4	5	4	23

5. Conclusions

Effective threat assessment is critical in military operations, where accurate identification and analysis of potential risks are essential for mission success and personnel safety. Segmentation and labelling provide a structured approach to understanding complex scenarios and identifying threats accurately. However, data scarcity in military domains poses a challenge. Integrating labelling and segmentation with generative learning models addresses this challenge by allowing for the creation of synthetic data, enabling realistic simulations of various threats. This approach enhances training and preparation for diverse scenarios while continuously improving threat assessment capabilities. By leveraging advanced methods, military organizations are enabled to better identify and respond to evolving threats, ensuring operational readiness and personnel safety.

Acknowledgements

The authors received support from the FaRADAI project (ref. 101103386) funded by the European Commission under the European Defence Fund.

References

Amico, V., Bruzzone, A. G., & Guha, R. (2000, July). Critical issues in simulation. In Summer Computer Simulation Conference (pp. 893-898). Society for Computer Simulation International; 1998.

Bang, M. Liwång, H. (2016). Influences on threat assessment in a military context. Defense & Security Analysis, 32(3), 264-277.

Blasch, E., Pham, T., Chong, C. Y., Koch, W., Leung, H., Braines, D., & Abdelzaher, T. (2021). Machine learning/artificial intelligence for sensor data fusion–opportunities and challenges. IEEE Aerospace and Electronic Systems Magazine, 36(7), 80-93.

Bruzzone, A., Remondino, M., Battista, U., Tardito, G., & Santoni, F. T. (2021). Modelling & Data Fusion to support Acquisition in Defence.

Bruzzone, A. G., Di Matteo, R., & Sinelshchikov, K. (2018). Strategic Engineering & Innovative Modeling Paradigms. In Workshop on Applied Modelling & Simulation (p. 14).

Bruzzone, A. G. & Massei, M. (2017). Simulation-based military training. Guide to Simulation-Based Disciplines: Advancing Our Computational Future, 315-361.

Bruzzone, A. G., Corso, M., Longo, F., Massei, M., & Tremori, A. (2014). Data Fusion and simulation as decision support system in naval operations. Proc. of HMS, Bordeaux, France, September.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587)..

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In Proc. of the IEEE international conference on computer vision (pp. 2961-2969).

Hussain, S., Kamal, A., Ahmad, S., Rasool, G., & Iqbal, S. (2014). Threat modelling methodologies: a survey. Sci. Int.(Lahore), 26(4), 1607-1609.

Liang, M., & Hu, X. (2015). Recurrent convolutional neural network for object recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3367-3375).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

Sizintsev, M., Rajvanshi, A., Chiu, H. P., Kaighn, K., Samarasekera, S., & Snyder, D. P. (2019, September). Multi-sensor fusion for motion estimation in visually-degraded environments. In 2019 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR) (pp. 7-14). IEEE.

Vakil, A., Liu, J., Zulch, P., Blasch, E., Ewing, R., & Li, J. (2021). A survey of multimodal sensor fusion for passive RF & EO information integration. IEEE Aerospace and Electronic Systems Magazine, 36(7), 44-61.

Waltz, E., & Llinas, J. (1990). Multisensor data fusion (Vol. 685). Boston: Artech house.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016). CNN-RNN: A unified framework for multi-label image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2285-2294).

Wei, Y., Xia, W., Lin, M., Huang, J., Ni, B., Dong, J., ... & Yan, S. (2015). HCP: A flexible CNN framework for multi-label image classification. IEEE transactions on pattern analysis and machine intelligence, 38(9), 1901-1907.

Wu, M., Yue, H., Wang, J., Huang, Y., Liu, M., Jiang, Y., ... & Zeng, C. (2020). Object detection based on RGC mask R-CNN. IET Image Processing, 14(8), 1502-1508.

Zhang, Y. J. (2001, August). A review of recent evaluation methods for image segmentation. In Proceedings of the sixth international symposium on signal processing and its applications (Cat. No. 01EX467) (Vol. 1, pp. 148-151). IEEE.

Zhou, C., Loy, C. C., & Dai, B. (2022, October). Extract free dense labels from clip. In European Conference on Computer Vision (pp. 696-712). Cham: Springer Nature Switzerland.