Data augmentation for UAV-captured vessel images in maritime surveillance using multimodal language and diffusion models

Le Thi Thu Hong, Pham Thu Huong, Doan Quang Tu, Nguyen Chi Thanh*

Institute of Information Technology and Electronics, Academy of Military Science and Technology, 17 Hoang Sam, Nghia Do, Hanoi, Vietnam.
*Corresponding author: thanhnc@ioit.ai.vn
Received 28 Jul. 2025; Revised 24 Sep. 2025; Accepted 10 Oct. 2025; Published 30 Oct. 2025.
DOI: https://doi.org/10.54939/1859-1043.j.mst.IITE.2025.160-168

ABSTRACT

In maritime surveillance, UAV-based vessel detection is essential for ensuring security and safety at sea. However, limited and non-diverse annotated data often restrict model performance in complex maritime environments. This study introduces a novel data augmentation pipeline using multimodal generative models to enhance training datasets with realistic synthetic images. Scene descriptions are automatically generated from UAV imagery using Gemma, a lightweight multimodal language model, and then used to guide FLUX, a text-to-image diffusion model, in creating diverse vessel-centric scenes under varying environmental conditions. A hybrid annotation strategy combines YOLO-World for initial object proposals with manual refinement to ensure label accuracy. The augmented dataset is integrated with the original data to train a vessel detection model. Experiments on the VESSELImg benchmark demonstrate that the proposed approach improves the YOLOv11 detector's mean average precision (mAP) from 0.775 to 0.805 at IoU thresholds of 0.50:0.95. These results validate the effectiveness of combining multimodal diffusion and language models for domain-specific data synthesis, offering improved generalization and robustness in UAV-based maritime vessel detection.

Keywords: Diffusion; Image synthesis; Data augmentation; Vessel detection.

1. INTRODUCTION

The surveillance and monitoring of vessels are vital for maintaining maritime security and safety. Accurate detection, identification, and tracking of vessels are essential for mitigating threats, preventing illegal activities such as smuggling or unauthorized entry, and ensuring maritime order. Recent technological advancements, particularly the use of unmanned aerial vehicles (UAVs), have greatly enhanced maritime observation capabilities. UAV-based imaging systems provide high-resolution, real-time data that overcome many limitations of conventional monitoring methods [1]. Figure 1 illustrates examples of vessel imagery captured by UAVs operating at sea. Meanwhile, advancements in artificial intelligence (AI) and computer vision have enabled machines to interpret visual data with unprecedented precision. Within this context, vessel detection from UAV imagery has become a crucial application, supporting effective maritime surveillance. However, the success of such detection models relies heavily on large, diverse, and high-quality training datasets - resources that remain difficult and costly to obtain in real-world maritime conditions. Data augmentation techniques [2] have become an effective strategy to overcome the limitations posed by insufficient annotated training data. By generating new samples through geometric transformations such as rotation, translation, and scaling, these methods improve dataset diversity and model robustness. However, conventional augmentation approaches often fail to represent the complex variability of maritime environments, including dynamic weather, illumination changes, time of capture, and regional sea characteristics. Recent advancements in generative artificial intelligence (AI), particularly Generative Adversarial Networks (GANs) and diffusion models [3], have significantly improved image synthesis. Diffusion models, in particular, excel at learning data distributions to generate diverse, high-

fidelity images, often outperforming traditional generative methods. Moreover, their ability to produce semantically meaningful images from textual prompts enables integration with multimodal language models, forming a novel augmentation paradigm that effectively enhances diversity and realism in domain-specific datasets such as maritime imagery.

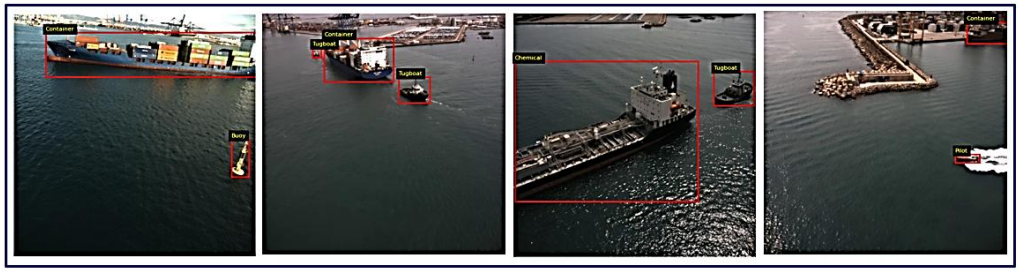

Figure 1. *Examples of vessel images captured by UAV from VESSELImg dataset.*

In this study, we present a novel data augmentation framework for UAV-captured vessel imagery that leverages multimodal generative models to enhance training datasets with high-quality synthetic samples. Scene descriptions are first generated from real UAV vessel images using Gemma [4], a lightweight multimodal language model by Google DeepMind. These textual descriptions are then expanded with variations in weather, illumination, time of day, and viewpoint to create prompts for FLUX [5], a text-to-image diffusion model. The synthesized images are annotated using a hybrid labeling process - initial object proposals from YOLO-World [6] are refined through manual verification. The resulting synthetic data are merged with original samples to train the YOLOv11 [7] vessel detector, improving detection accuracy. Our key contributions are summarized as follows: (1) a multimodal augmentation pipeline integrating language and diffusion models; (2) a prompt diversification strategy; and (3) a hybrid annotation method ensuring precise labeling.

The structure of this paper is as follows. Section II briefly reviews Diffusion-based methods for augmenting UAV-based image dataset. Section III details our proposed data augmentation approach using a text-to-image diffusion model to enhance ship detection. Section IV presents the experimental setup. The results and discussion are presented in section V, followed by conclusions in section VI.

2. RELATED WORK

Data augmentation is essential for enhancing the performance and generalization of deep learning models, especially in remote sensing and UAV imagery, where labeled data is limited and costly. Conventional techniques such as flipping, rotation, scaling, and color jittering [8] have been extensively applied to expand datasets synthetically. However, these methods provide only low-level variations and fail to introduce meaningful semantic diversity or new visual content. In UAV-based vessel detection, where objects exhibit wide variability in size, shape, orientation, and environmental context, traditional augmentation remains insufficient to capture complex real-world conditions.

In recent years, generative models have become powerful approaches for data augmentation, enabling the synthesis of novel and realistic samples that improve dataset diversity. Among them, GANs have attracted considerable attention for their ability to model complex data distributions and generate high-fidelity images. A GAN consists of two competing networks - a generator that produces synthetic data and a discriminator that assesses its authenticity - trained in an adversarial framework. This process enables GANs to create diverse and detailed imagery beyond traditional augmentation methods. Consequently, GANs have been widely adopted for synthetic image generation in medical imaging, remote sensing, and maritime surveillance [9, 10]. For example,

Zhang et al. utilized conditional GANs to enrich aerial scene datasets, achieving improved classification performance in imbalanced cases [11]. However, GANs remain challenging to train, prone to instability and mode collapse, and often struggle to capture complex structural variations crucial for vessel detection.

Diffusion models have recently emerged as a stable and scalable alternative to GANs for generating high-quality images [12]. These models operate by progressively denoising Gaussian noise through a diffusion process. The Denoising Diffusion Probabilistic Model (DDPM), introduced by Ho et al. [13], achieved state-of-the-art results across multiple image generation benchmarks. Building on this foundation, Saharia et al. developed conditional diffusion models that enable task-specific or class-conditional synthesis [14], offering fine-grained control essential in structured domains such as vessel detection. Although diffusion-based augmentation for UAV-based vessel detection remains underexplored, it shows strong potential. Studies in medical imaging and satellite observation demonstrate that diffusion models can generate diverse, realistic data that enhance model performance on small or imbalanced datasets [15]. Their capacity to produce semantically rich, high-fidelity samples makes them highly suitable for improving dataset diversity and robustness in complex maritime environments.

3. PROPOSED METHOD

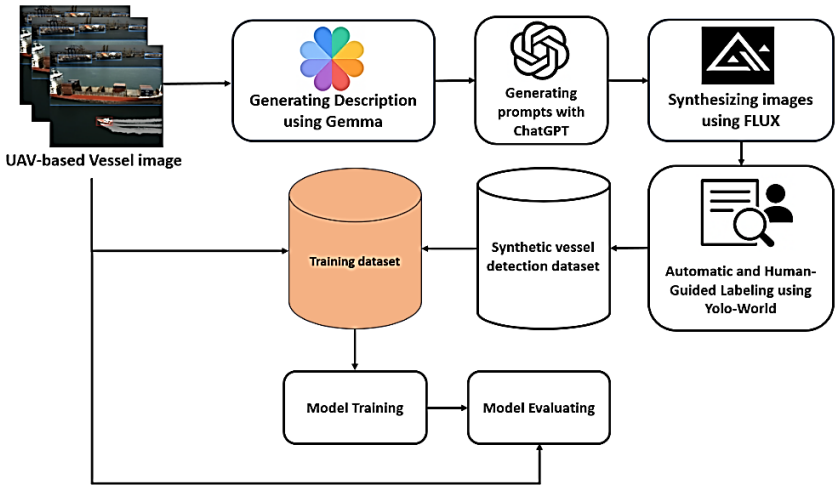

Figure 2. *Framework for UAV-Based vessel detection dataset augmentation.*

Figure 2 shows our proposed framework for UAV-Based vessel detection dataset augmentation. This framework is designed to enhance vessel detection capabilities by augmenting UAV-based datasets with synthetically generated data. The process follows these key steps:

- *UAV-Based vessel image collection*: The pipeline begins with the collection of real vessel images captured by UAVs.

- *Generating descriptions using gemma*: Descriptive captions are generated for the collected images using the multimodal language model Gemma.

-*Prompt generation with ChatGPT*: The image descriptions are used by ChatGPT to create diverse prompts for generating new synthetic images.

- *Image synthesis with FLUX*: These prompts are then processed by the text-to-image diffusion model FLUX to create new vessel images that match the given prompts.

- *Automatic and human-guided labeling*: YOLO-World is used to automatically label synthetic images. Human annotators then refine labels, forming a hybrid workflow that ensures both efficiency and accuracy.

- *Synthetic dataset creation*: The labeled synthetic images are compiled into a synthetic vessel detection dataset.

- *Merging with real data*: The synthetic dataset is merged with the original UAV-based images to form an enriched training dataset.

- *Model training:* The combined dataset is used to train Yolov11 vessel detection model.

- *Model evaluation:* The performance of the trained model is evaluated using appropriate evaluation methods to assess the effectiveness of the data augmentation approach.

4. EXPERIMENTAL SETUP

4.1. Dataset

We conducted experiments on the VESSELImg benchmark dataset [15], a large UAV-based vessel image dataset for sea surveillance. The dataset is annotated for object detection with six labeled object classes, it is divided into training, validation, and test sets. Table 1 provides statistics on the number of images and objects in this dataset.

Table 1. *Statistics on the number of images and objects in VESSELImg dataset.*

	Num of files	**Num of objects**	**Container**	**Tugboat**	**Passenger-RoRo**	**Chemical**	**Buoy**	**Pilot**
Train	4662	9008	3901	3029	1163	224	391	300
Valid	1222	2518	1102	913	361	59	113	70
Test	608	1253	532	422	179	24	69	27
Total	**6492**	**12779**	**5535**	**4364**	**1703**	**307**	**573**	**397**

4.2. Implementation

For image caption generation, we employed the '*google/gemma-3-4b-it*' model [16], and for image synthesis, we utilized the '*black-forest-labs/FLUX.1-dev*' model [17]. Both models were obtained from the Hugging Face repository. Additionally, the object detection models, YOLO-World and YOLO11, were sourced from Ultralytics. All experiments are performed on AMST AI Platform (AAP). It provides environments to train and deploy ML models easily. Regarding evaluation metrics, we utilize a common mAP across the methods, encompassing the size types of the mAP: average $mAP, mAP@50, mAP@75, mAP@75.

5. RESULTS AND DISSCUSSION

5.1. Results of UAV-based vessel image synthesis

In our experiments, we leveraged FLUX to generate realistic synthetic images of ships in maritime environments, aiming to enrich the training dataset with diverse real-world scenarios. We first used Gemma to create realistic scene descriptions from the VEESELImage dataset. To further enhance data diversity, ChatGPT was employed to generate context-aware prompts based on these descriptions, guiding FLUX in synthesizing UAV-based vessel images. The generated scenes covered a wide range of weather conditions, lighting variations, times of day, and camera angles. Figure 3 illustrates a real image from the VEESEL dataset, along with the simulated description from Gemma and the derived prompts. Figure 4 shows examples of prompts and the corresponding FLUX-generated images, which demonstrate significant diversity and realism compared to the original input.

After generating the synthetic images, we employed YOLO-World to automatically detect and annotate vessel objects, classifying them into six categories: *Container Ship, Chemical/Bulk Carrier, Passenger/Ro-Ro Vessel, Tugboat, and Pilot Vessel*. To ensure high annotation quality, human reviewers refined the labels and removed irrelevant samples. The final synthetic vessel detection dataset includes 6,444 annotated images with 15,120 vessel instances: 6,447 Containers,

4,143 Tugboats, 3,174 Passenger/Ro-Ro Vessels, 1,045 Pilots, 160 Buoys, and 142 Chemical/Bulk Carriers, providing diverse object types for robust vessel detection training. We refer to this dataset as "*FluxSynVESSELImg*", and utilize it to augment the training data for vessel detection models.

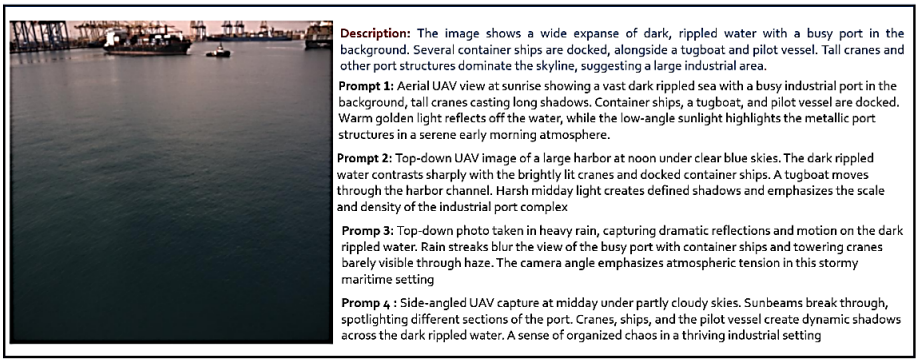

Figure 3. *An example image from the VESSELImg dataset, along with the description generated by Gemma and the prompts derived from that description.*

Figure 4. *Prompts and the corresponding images generated from each prompt.*

5.2. Results of vessel detection

We first trained YOLOv11 models for vessel detection using the real-world VESSELImg dataset. Subsequently, synthetic images from "*FluxSynVESSELImg*" dataset were incrementally added for training. Model performance was evaluated on the VESSELImg test set. Table 2 reports detection results across models trained with varying amounts of synthetic data for all classes, while table 3 presents class-wise results. The results indicate that incorporating synthetic data from FluxSynVESSELImg generally enhances YOLOv11's performance, particularly at moderate levels of augmentation. Table 2 shows a peak in overall mAP@50-95 when using 4283 synthetic images, increasing the score from 0.775 to 0.807, which suggests an optimal balance between real and synthetic data. In table 3, class-wise performance also improves, especially for underrepresented classes such as 'Pilot' and 'Chemical', which benefit from better recall and mAP@50-95 with additional synthetic samples. However, excessive augmentation (e.g., 6444 images) slightly reduces performance in some classes, indicating possible domain shift or overfitting to synthetic patterns.

5.3. Discussions

This study proposes a novel data augmentation framework for UAV-captured vessel imagery by leveraging a multimodal language model and diffusion-based image synthesis, combined with a hybrid labeling strategy. Experimental results on the real-world UAV dataset VESSELImg demonstrate improved detection accuracy, particularly for rare and underrepresented vessel classes. Beyond vessel detection, the framework is adaptable to various object types such as warships, submarines, tanks, missiles, vehicles, and buildings by modifying prompt semantics and domain-specific attributes. This generalizable approach offers a data-efficient solution for object detection in complex, low-data environments, supporting broader integration of generative AI tools into UAV-based image analysis pipelines.

Table 2. *Performance of YOLO11 on the test set across all classes.*

Training data	Number of files	Number of objects	Precision	Recall	mAP @50	mAP @50-95
VESSELImg	*4,262*	*9008*	***0.968***	*0.933*	*0.977*	*0.775*
VESSELImg+708FluxSyn	4,970	10596	0.946	0.954	0.975	0.766
VESSELImg+2146FluxSyn	6,408	14011	0.953	0.931	0.971	0.769
VESSELImg+4283FluxSyn	8,545	18993	0.935	**0.952**	**0.979**	**0.807**
VESSELImg+6444FluxSyn	10,706	21428	0.934	0.948	0.976	0.772

Table 3. *Performance of YOLO11 on the test set for each class.*

Training data	Num of object	Precision	Recall	mAP@50	mAP @50-95
Container					
VESSELImg	*3901*	***0.95***	*0.966*	*0.989*	*0.873*
VESSELImg+708 FluxSynVESSELImg	4584	0.908	0.983	0.981	0.852
VESSELImg+2146 FluxSynVESSELImg	5952	0.941	0.963	0.984	0.866
VESSELImg+4283 FluxSynVESSELImg	8154	0.889	**0.987**	**0.989**	**0.875**
VESSELImg+6444 FluxSynVESSELImg	10348	0.886	0.991	0.989	0.858

Training data	Num of object	Precision	Recall	mAP@50	mAP @50-95
Passenger-RoRo					
VESSELImg	*1163*	***0.97***	*0.961*	*0.989*	*0.858*
VESSELImg+708 FluxSynVESSELImg	1430	0.932	0.972	0.988	0.84
VESSELImg+2146 FluxSynVESSELImg	2234	0.928	0.94	0.981	0.84
VESSELImg+4283 FluxSynVESSELImg	3233	0.868	**0.994**	**0.989**	**0.866**
VESSELImg+6444 FluxSynVESSELImg	4337	0.87	0.983	0.985	0.829
Tugboat					
VESSELImg	*3029*	*0.991*	*0.936*	*0.992*	*0.79*
VESSELImg+708 FluxSynVESSELImg	3497	0.99	0.946	0.993	0.772
VESSELImg+2146 FluxSynVESSELImg	4409	0.988	0.948	0.988	0.771
VESSELImg+4283 FluxSynVESSELImg	5775	0.972	0.979	0.992	**0.786**
VESSELImg+6444 FluxSynVESSELImg	7172	0.976	0.969	0.988	0.768
Pilot					
VESSELImg	*300*	0.94	0.778	0.913	0.534
VESSELImg+708 FluxSynVESSELImg	391	0.92	0.85	0.914	0.472
VESSELImg+2146 FluxSynVESSELImg	670	0.93	0.778	0.898	0.472
VESSELImg+4283 FluxSynVESSELImg	1005	0.956	0.798	0.924	**0.542**
VESSELImg+6444 FluxSynVESSELImg	1345	0.955	0.79	0.925	0.502
Chemical					
VESSELImg	*224*	*0.957*	*1*	*0.993*	*0.92*
VESSELImg+708 FluxSynVESSELImg	257	0.955	1	0.995	0.923
VESSELImg+2146 FluxSynVESSELImg	274	0.945	1	0.995	0.938
VESSELImg+4283 FluxSynVESSELImg	322	0.948	1	0.995	**0.967**
VESSELImg+6444 FluxSynVESSELImg	366	0.944	1	0.993	0.929
Buoy					
VESSELImg	*391*	*0.999*	*0.957*	*0.985*	*0.734*
VESSELImg+708 FluxSynVESSELImg	437	0.969	0.971	0.978	0.738
VESSELImg+2146 FluxSynVESSELImg	472	0.989	0.957	0.977	0.729
VESSELImg+4283 FluxSynVESSELImg	504	0.974	0.957	0.982	**0.803**
VESSELImg+6444 FluxSynVESSELImg	551	0.974	0.957	0.978	0.747

6. CONCLUSIONS

In this study, we propose a novel data augmentation workflow for training object detection models on UAV-acquired vessel imagery by leveraging multimodal language models (LLMs) and text-to-image diffusion models. Our approach enriches training datasets by generating diverse

variations in environmental conditions, weather, time of day, and camera perspectives. Through the use of natural language prompts with varying semantics, we synthesize realistic and context-aware images that help improve model robustness. We evaluate the proposed augmentation method using the YOLOv11 detector, demonstrating significant improvements in detection performance and generalization on the real-world UAV dataset VESSELImg. These results highlight the effectiveness of integrating LLM-guided prompt generation with diffusion-based synthesis to increase training data diversity and enhance object detection accuracy in UAV imagery. However, the current study is limited to a single UAV dataset (VESSELImg) and one object detection model (YOLOv11). In future work, we plan to evaluate the proposed framework across additional UAV-acquired datasets and a broader range of detection architectures. These extensions aim to further validate the adaptability, scalability, and reliability of the proposed approach for object detection in diverse aerial imaging scenarios.

Acknowledgment: *This work was supported by the national-level scientific project KC-4.0-52/19-30 under the KC-4.0/19-30 program, Ministry of Science and Technology, Vietnam.*

REFERENCES

[1]. Cheng, S., Zhu, Y., & Wu, S. *"Deep learning based efficient ship detection from drone-captured images for maritime surveillance."* Ocean engineering, 285, 115440, (2023).

[2]. Shorten, C., & Khoshgoftaar, T. M. *"A survey on image data augmentation for deep learning."* Journal of big data, 6(1), 1–48, (2019).

[3]. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. *"High-resolution image synthesis with latent diffusion models."* Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695, (2022).

[4]. Team, G et al. *"Gemma: Open models based on gemini research and technology."* arXiv preprint arXiv:2403.08295, (2024).

[5]. Black Forest Lab. *"FLUX."*, (2024). https://github.com/black-forest-labs/flux.

[6]. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., & Shan, Y. *"Yolo-world: Real-time open-vocabulary object detection."* Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, –, 16901–16911, (2024).

[7]. Glenn, J., & Jing, Q. *"Ultralytics YOLO11."*, (2024). https://github.com/ultralytics/ultralytics.

[8]. Goodfellow. I et al. *"Generative adversarial nets."* Advances in neural information processing systems, pp. 2672–2680, (2014).

[9]. Xu, M., Xie, L., Liu, Y., Wang, S., & Zhang, Y. *"Generative adversarial networks in remote sensing: A review."* ISPRS journal of photogrammetry and remote sensing, 166, 296–312, (2020).

[10]. Zhang, Y., Zhang, C., Zhang, Q., & Xie, W. *"Data augmentation with conditional GAN for aerial scene classification."* Remote sensing, 11(3), 243, (2019).

[11]. Dhariwal, P., & Nichol, A. *"Diffusion models beat GANs on image synthesis."* Advances in neural information processing systems, 34, 8780–8794, (2021).

[12]. Ho, J., Jain, A., & Abbeel, P. *"Denoising diffusion probabilistic models."* arXiv preprint arXiv:2006.11239, (2020).

[13]. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Salimans, T., Ho, J., Fleet, D., & Norouzi, M. *"Imagen: Text-to-image diffusion models."* International conference on machine learning (ICML), (2022).

[14]. Wolleb, J., Dejakum, K., Sandkühler, P., Reich, M., Lunz, S., & Cattin, P. C. *"Diffusion models for medical anomaly detection."* Medical image analysis, 76, 102327, (2022).

[15]. Rubis, B., Cacace, J., Rodriguez, J., Company, R., Tanner, M., Arzo, R., & Cayero, J. *"VESSELImg: A large UAV-based vessel image dataset for port surveillance."* International conference on unmanned aircraft systems (ICUAS), 76–83, (2024).

[16]. https://huggingface.co/google/gemma-3-4b-it

[17]. https://huggingface.co/black-forest-labs/FLUX.1-dev

TÓM TẮT

Tăng cường dữ liệu ảnh tàu thuyền chụp từ UAV trong giám sát hàng hải sử dụng mô hình ngôn ngữ đa phương thức và mô hình khuếch tán

Trong lĩnh vực giám sát hàng hải, việc phát hiện tàu thuyền từ ảnh chụp bởi thiết bị bay không người lái (UAV) đóng vai trò quan trọng trong đảm bảo an ninh và an toàn trên biển. Tuy nhiên, sự hạn chế về số lượng và tính đa dạng của dữ liệu gán nhãn thường làm giảm hiệu suất của các mô hình trong môi trường hàng hải phức tạp. Nghiên cứu này giới thiệu một quy trình tăng cường dữ liệu mới, sử dụng các mô hình sinh đa phương thức để tạo ra các mẫu tổng hợp chân thực nhằm mở rộng tập huấn luyện. Mô tả cảnh được tự động sinh từ ảnh UAV bằng Gemma, một mô hình ngôn ngữ đa phương thức gọn nhẹ, sau đó được dùng để hướng dẫn FLUX, một mô hình khuếch tán chuyển văn bản thành hình ảnh, tạo ra các cảnh có tàu trong nhiều điều kiện môi trường khác nhau. Chiến lược gán nhãn lại được áp dụng, kết hợp giữa dự đoán ban đầu của YOLO-World và tinh chỉnh thủ công nhằm đảm bảo độ chính xác của nhãn. Tập dữ liệu tổng hợp sau đó được kết hợp với dữ liệu gốc để huấn luyện mô hình phát hiện tàu. Thí nghiệm trên bộ dữ liệu VESSELImg cho thấy phương pháp đề xuất giúp mô hình YOLOv11 tăng chỉ số mAP từ 0.775 lên 0.805 ở ngưỡng IoU 0.50:0.95. Kết quả này khẳng định hiệu quả của việc tích hợp mô hình ngôn ngữ và khuếch tán đa phương thức trong tăng cường dữ liệu chuyên biệt, giúp cải thiện khả năng khái quát và độ bền vững của hệ thống phát hiện tàu từ UAV.

Từ khóa: Khuếch tán; Tổng hợp ảnh; Tăng cường dữ liệu; Phát hiện tàu thuyền.