Expert Systems With Applications 287 (2025) 128169

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Customization of the text-to-image diffusion model by fine-tuning for the generation of synthetic images of cyanobacterial blooms in lentic water bodies

Fredy Barrientos-Espillco [a,*], Gonzalo Pajares [b], José A. López-Orozco [a], Eva Besada-Portas [a]

[a] Department of Computer Architecture and Automation, University Complutense of Madrid 28040 Madrid, Spain
[b] Institute for Knowledge Technology, University Complutense of Madrid 28040 Madrid, Spain

ARTICLE INFO

Keywords:
Artificial intelligence
Machine vision systems
Text-to-image generation
Fine-tuning
Large language model
Convolutional neural networks
Autonomous surface vehicles
Cyanobacterial blooms

ABSTRACT

Cyanobacterial blooms emerge unpredictably on the surface of lentic water bodies, posing both ecological threats and public health risks. To effectively monitor these events, this study introduces the use of Machine Vision Systems (MVS) integrated into Autonomous Surface Vehicles (ASVs). These ASVs are capable of autonomous and safe navigation, enabling them to detect cyanobacterial blooms while avoiding obstacles. Convolutional Neural Networks (CNNs) are employed for early detection and continuous monitoring, but their effectiveness hinges on access to large, high-quality training datasets. Due to the sporadic and uncontrollable nature of bloom occurrences, acquiring sufficient real-world images for training and validating CNN models is a significant challenge. To overcome this, the Stable Diffusion XL (SDXL) text-to-image generative model is utilized to produce realistic synthetic images, ensuring a sufficient dataset for training. However, SDXL alone struggles to accurately depict cyanobacterial blooms. To address this limitation, DreamBooth is used to fine-tune SDXL with a small set of real bloom-specific image patches. To ensure the diversity of the synthetic dataset, detailed prompts for SDXL are generated using a Large Language Model (LLM). The combination of SDXL fine-tuning with LLM-driven prompts design applied to environmental monitoring and autonomous navigation in lentic environments represents the core innovation of this work. A dual-task CNN model is then trained on the synthetic dataset to simultaneously detect blooms and obstacles. Experimental results demonstrate the effectiveness and novelty of the proposed approach, showing improvements of up to 15.74% in object detection and 6.48% in semantic segmentation compared to the baseline dataset.

1. Introduction

Water is a fundamental natural resource, critical to maintaining life and supporting ecological balance on Earth. Despite its essential role, it is highly susceptible to various environmental stressors, including flooding, prolonged droughts, and contamination from hazardous pollutants. In recent years, cyanobacterial blooms have emerged as a significant global concern, increasingly observed across diverse aquatic systems worldwide and posing serious risks to environmental and public health. This proliferation constitutes not only a threat to human health, animals, and aquatic ecosystems (Graham et al., 2016; Huisman et al., 2018), but also leads to economic damage (Hamilton et al., 2014).

Early detection and continuous monitoring of cyanobacterial blooms are critical for effective water quality management. The conventional methodology primarily involves the manual collection of water samples followed by laboratory-based analyses, which are both resources-intensive and economically unfeasible for comprehensive coverage across all water bodies. Consequently, there is a compelling need to develop cost-effective and scalable monitoring solutions. Although remote sensing techniques utilizing satellite imagery (Ahn et al., 2006; Cannizzaro et al., 2019; Chen et al., 2020; Hu, 2009; Kutser et al., 2006) offer a non-invasive alternative, their efficacy is limited by meteorological constraints and the temporal resolution of satellite overpasses, rendering them inadequate for capturing the rapid temporal dynamics characteristic of cyanobacterial bloom events.

The recent study by Barrientos-Espillco et al. (2023) addresses the

* Corresponding author.
E-mail addresses: fredybar@ucm.es (F. Barrientos-Espillco), pajares@ucm.es (G. Pajares), jalopez@ucm.es (J.A. López-Orozco), ebesada@ucm.es (E. Besada-Portas).

https://doi.org/10.1016/j.eswa.2025.128169
Received 20 December 2024; Received in revised form 30 April 2025; Accepted 7 May 2025
Available online 15 May 2025

detection of cyanobacterial blooms using Autonomous Surface Vehicles (ASVs) equipped with Machine Vision Systems (MVS). As noted in Barrientos-Espillco et al. (2023), these blooms proliferate in lentic or low-flow aquatic environments —such as lakes, lagoons, reservoirs, and wetlands— and require continuous monitoring to mitigate their potential adverse impacts. Thus, ASVs navigate with the support of the MVS, enabling them to detect cyanobacterial blooms within these aquatic ecosystems while ensuring safe navigation and effective obstacle avoidance. The method in Barrientos-Espillco et al. (2023) uses techniques based on Deep Learning (DL) using synthetic images.

The use of synthetic images is forced by the absence of sufficient real images to train and validate DL models containing cyanobacterial blooms. Moreover, the few real ones that exist present licensing problems and the synthetic ones used in Barrientos-Espillco et al. (2023) do not reach a sufficient level of realism, mainly due to the fusion between background and foreground as well as an imbalance in the class categories. On the other hand, capturing real images is a difficult challenge due to the dynamic behavior of blooms, caused by factors such as biological growth, competition with other organisms, vertical self-moving ability in the water column, and three-dimensional displacement due to water currents and wind (Besada-Portas et al., 2023). To overcome these limitations, we propose a new novel approach to generate enough synthetic and realistic images, so that DL models can be trained and validated accordingly. These images must contain elements that can be considered navigational obstacles for ASVs, such as submerged rocks, floating trunks, aquatic vegetation, swimmers, canoes, birds, among others. In addition, it is crucial that the synthetic images include cyanobacterial blooms in the water body, as shown in Fig. 1. This is the starting hypothesis.

Recent advances in text-to-image models, such as the models, DALL-E2 (Ramesh et al., 2022), Image (Saharia et al., 2022), and Stable Diffusion (Rombach et al., 2022), have provoked a great deal of attention for their ability to generate realistic images from natural language prompts. Despite its potential, the main limitation of these models lies in their inability to synthesize novel concepts from inputs prompts, in particular cyanobacterial blooms, which is our main goal. This inability stems from the absence of these concepts in the datasets used to train these models. In the face of this limitation, recent studies such as Dong et al. (2023), Gal et al., (2022) and Ruiz et al. (2023) have proposed personalized image generation techniques based on the fine-tuning of previously trained text-to-image models using only reduced sets of images representative of specific concepts.

In this context, and in order to address the aforementioned limitations, our approach is based on the use of the DreamBooth method (Ruiz et al., 2023). Using this technique, we perform a fine-tuning of the Stable Diffusion XL (SDXL) text-to-image diffusion model (Podell et al., 2023) using a limited collection of real image patches depicting cyanobacterial blooms, each linked to a unique identifier that helps the model learn and reproduce specific visual concepts. This technique enables the creation of synthetic yet realistic images that capture the complexity of lentic aquatic environments, helping to overcome data scarcity challenges in the training and validation of DL models for autonomous aquatic navigation and bloom detection.

To further diversify the dataset (capturing variations in bloom appearance, location, and environmental conditions) prompts to feed SDXL are required. The manual design of these prompts represents a laborious task, due to the need to create an extensive and diversified dataset of lentic water images that meet the aforementioned characteristics. In this context, for the generation of natural language prompts with sufficient variability to satisfy the requirements of this study, we use the Large Language Model (LLM) called LLaMa 2 (Touvron et al., 2023), recently developed by Meta.

To validate the feasibility of this approach, our study uses the dual-task architecture proposed in Barrientos-Espillco et al. (2024), which simultaneously addresses the tasks of object detection and semantic segmentation based on Convolutional Neural Networks (CNNs). This framework allows the accurate identification and localization of both navigational obstacles and amorphous semantic structures, such as cyanobacterial blooms from a MSV on board an ASV. The approach proposed in Barrientos-Espillco et al. (2024) sufficiently describes the mathematical part that supports the architecture proposed here.

In summary, this study focuses on generating high-fidelity synthetic images that depict both cyanobacterial blooms and navigational obstacles in lentic water bodies. To achieve this, we apply fine-tuning of the SDXL text-to-image diffusion model using the DreamBooth technique and a small set of real cyanobacterial bloom patches. A key element in preserving the diversity of the synthetic dataset is the use of a large language model (LLM), specifically LLaMa 2, to create natural language prompts that guide the SDXL model. The primary contribution and innovation of our work lies in integrating SDXL fine-tuning with LLM-driven prompts generation for synthetic data creation and applying this strategy to environmental monitoring and autonomous navigation in lentic ecosystems. This is broken down into the following five key points:

- We propose a novel approach for the synthesis of full-scene lentic water images using the DreamBooth technique to fine-tune the SDXL model and learn the intrinsic features of real cyanobacterial patches.
- We generate prompts in natural language with high variability using the LLM, in particular LLaMa 2.
- We automatically annotate the synthetic images using the Grounding DINO model (Liu et al., 2023), which generates bounding boxes around the objects present in the scene. These bounding boxes are used as input to the Segment Anything Model (SAM) (Kirillov et al., 2023), which converts them into segmentation masks corresponding to each object in the image.

Fig. 1. Synthetic images of cyanobacterial blooms in lentic water bodies generated by the custom Stable Diffusion XL text-to-image model. Input images of a concept (left). Generated images of the concept under various adverse environmental conditions, using different prompts (right).

- We use generated and annotated synthetic images to train the dual-task model based on CNNs that integrates object detection and semantic segmentation.
- We perform a comparative analysis of the performance of the dual-task model trained independently on three different datasets: synthetic images generated using the SDXL custom model, the dataset presented in study by Barrientos-Espillco et al. (2024), and the combination of both datasets (synthetic images and the study dataset Barrientos-Espillco et al. (2024)).

Beyond aquatic environmental monitoring, the proposed synthetic image generation strategy has potential applications in other domains where access to real data is limited. For example, in medical imaging, diffusion models have been applied to generate synthetic cancer images to support diagnostic models while preserving patient privacy (Kidder, 2024). In the field of remote sensing, text-to-image diffusion models have been used to generate synthetic satellite images for land cover analysis and environmental change detection in data-scarce regions (Nguyen et al., 2024). Additionally, our approach could be adapted for oil spill detection, identification of floating waste, or precision agriculture, where real-time visual monitoring is critical, but datasets are often limited. These examples demonstrate the versatility of our method for addressing data scarcity in multiple high-impact domains.

The rest of this paper is organized as follows. First, we review related work in Section 2. Then, in Section 3, we describe the method in detail. Next, in Section 4, we present experimental results and discussion. Finally, we draw our conclusions in Section 5.

2. Related work

2.1. Text-to-image synthesis

The challenge of creating images from textual descriptions, referred to as text-to-image synthesis, has gained growing interest within the fields of computer vision and natural language processing (Alhabeeb & Al-Shargabi, 2024; Bie et al., 2025). This capability holds considerable promise for a broad spectrum of innovative applications, from creative content generation to advanced design processes, among others (Bie et al., 2025). Text-to-image synthesis aims to bridge the intrinsic semantic gap between human language and visual content, offering a powerful means of translating abstract ideas expressed in natural language into concrete and interpretable visual representations.

In recent years, this field has experienced substantial advances, driven by the development of deep learning techniques, the increasing availability of large datasets, and wider access to high-performance computational resources. The evolution has been remarkable, moving from conventional computational methods to sophisticated deep learning-based models, such as Generative Adversarial Networks (GANs), autoregressive models, transformer networks and, more recently, diffusion models (Bie et al., 2025; Zhang & Tang, 2024).

The main goal of this line of research is to develop models capable of understanding and interpreting complex textual descriptions, generating realistic and coherent images that accurately reflect the semantic details, spatial relationships, styles and concepts present in the input text. Although GANs initially led developments in this field, the recent emergence of diffusion models has marked a major shift, owing to their

Table 1
Comparative analysis of pioneering Text-to-Image synthesis models.

Model Name	Organization	Architecture	Image Resolution	Main Contribution	Limitation
alignDRAW (Mansimov et al., 2016)	University of Toronto	Recurrent VAE with Attention	32x32 pixels	First deep learning-based text-to-image model; demonstrated generalization images from text captions.	Low resolution (32x32), limited to simple images, low diversity.
Generative Adversarial Text to Image Synthesis (Reed et al., 2016)	Multiple Institutions	Deep Convolutional GAN (DCGAN)	64x64 pixels	First to use GANs for text-to-image synthesis; generated plausible images of birds and flowers from descriptions.	Domain-specific (birds, flowers), low resolution (64x64), required separate training for different domains.
StackGAN (H. Zhang et al., 2017)	Researchers	Stacked GANs (Stage-I & Stage-II)	256x256 pixels	High-resolution images via sketch-refinement; Conditioning Augmentation	Unstable learning, mode collapse.
AttnGAN (Xu et al., 2018)	Microsoft Research	GAN with attention mechanisms	256x256 pixels	Introduced attention mechanisms for fine-grained text-to-image generation; improved image-text alignment.	Still GAN-based (potential mode collapse), resolution limited to 256x256.
PixelRNN (Oord et al., 2016)	DeepMind	Autoregressive RNN	Low resolution (e.g., 32x32)	Captures long-range dependencies, high-quality generation.	Slow training and generation.
GigaGAN (Kang et al., 2023)	CMU & Adobe Research	GAN	512x512 pixels	Scalable GAN for text-to-image, fast inference, latent space control.	Training instability, closed source.
DALL-E (Ramesh et al., 2021)	OpenAI	Autoregressive Transformer	256x256 pixels	First to use transformers for text-to-image synthesis at scale; generated diverse and creative images from text prompts.	Resolution limited to 256x256; some artifacts in complex images.
Muse (Chang et al., 2023)	Google	Transformer	1024x1024 pixels	Efficient parallel decoding; zero-shot image editing.	Struggles with long phrases and high object counts.
DDPM (Ho et al., 2020)	UC Berkeley	Diffusion Model	High quality on various resolutions	High-quality image synthesis, stable training, classifier guidance.	Slow sampling speed.
GLIDE (Nichol et al., 2022)	OpenAI	Diffusion Model	64x64 base, up to 256x256 pixels	Photorealistic generation, classifier-free guidance, inpainting.	Slower inference, struggled with shapes.
DALL-E 2 (Ramesh et al., 2022)	OpenAI	Diffusion model with CLIP guidance	1024x1024 pixels	Significantly improved image quality and resolution using diffusion models; incorporated inpainting and outpainting features.	Closed source; requires API access; risk of misuse.
Stable Diffusion (Rombach et al., 2022)	Stablity AI	Latent Diffusion Model	512x512 (v1.4), 1024x1024 (XL)	Open-source text-to-image model; widely adopted; allows for fine-tuning and customization. Versions include 1.4, 1.5, 2.0, 2.1, XL (1024x1024), and 3, each with improvements in quality and features.	Base models limited to 512x512 (requires upscaling for higher resolution); potential for generating harmful content if not moderated.
Image (Saharia et al., 2022)	Goolgle Brain	Diffusion model with large language model for text encoding	1024x1024 pixels	Achieved state-of-the-art text-to-image generation with high FID scores; uses a large pretrained language model for text encoding.	Not publicly available; research model only.

superior capability to produce high-quality, visually faithful images. (Alhabeeb & Al-Shargabi, 2024).

Table 1 provides a comparative summary of the leading models in text-to-image synthesis, categorized by their underlying architecture. It includes key information such as the name of the model, the organization(s) involved in its development, the architectural approach adopted, the image resolution achieved, the main contribution or innovation of the model, and the limitations identified in the specialized literature.

Overall, the early text-to-image synthesis models exhibit notable differences in their architecture, resolution capacity, key contributions, and inherent limitations. Nevertheless, they share a common objective: to interpret textual descriptions and produce realistic, coherent images. The following description summarizes the main features and distinctions of these models, organized by their underlying architecture:

GANs-based models. The initial GAN-based models, such as alignDRAW (Mansimov et al., 2016) and the model proposed by Reed et al. (2016) generated low resolution images (32x32 to 64x64 pixels). Subsequently, models such as StackGAN (Zhang et al., 2017) and AttnGAN (Xu et al., 2018) were able to increase the resolution up to 256x256 pixels using stacked architectures and attention mechanisms. More recently, models such as GigaGAN (Kang et al., 2023) have increased the resolution to 512x512 pixels with upscaling. The primary contribution of this family of models was their pioneering success in applying generative adversarial networks to text-to-image synthesis (Generative Adversarial Text to Image Synthesis), to introduce high resolution generation by progressive refinement (StackGAN), to incorporate word-level attention for fine details (AttnGAN). Despite advancements, GAN models have traditionally encountered considerable limitations, including training instability, mode collapse, and generally lower image quality compared to autoregressive or diffusion models, particularly when applied to large-scale datasets.

Autoregressive models. These models generate images sequentially, pixel by pixel, which allows them to capture long-range spatial dependencies. Early models like PixelRNN (van den Oord et al., 2016) exemplify this approach, though it comes with the trade-off of slower generation. Subsequent models, like DALL-E (Ramesh et al., 2021), use Transformer architectures to perform autoregressive modeling of image tokens, enabling zero-shot generation and the interpretation of complex prompts. However, a key limitation of autoregressive models is their slow generation speed, which results from the sequential nature of the processing.

Models based on non-autoregressive Transformers. To overcome the speed restrictions imposed by autoregressive models, new architectures have emerged allowing parallel decoding. Muse (Chang et al., 2023) is a notable example in this category, showcasing efficient image generation and zero-shot editing capabilities. Despite these advancements, these models still face challenges in interpreting lengthy sentences and accurately counting objects, which restricts their performance in more complex scenarios.

Diffusion Models. Diffusion models, starting with Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020), demonstrated superior ability to generate high quality and diverse images with more stable training compared to GANs. GLIDE (Nichol et al., 2022) explored different guidance methods (with and without classifier) to improve quality and text similarity. Stable Diffusion (Rombach et al., 2022) revolutionized the field by applying diffusion in latent space, drastically reducing computational costs and allowing its widespread use. The successive versions of this model (1.4, 1.5, 2.0, 2.1, XL, 3) reflect continuous advancements in resolution, quality, prompt interpretation, and computational efficiency. Nevertheless, limitations persist, including difficulty in accurately representing fine details (such as human anatomy and readable text), as well as interpreting complex prompts, and mitigating biases in the training data.

In summary, the progression of text-to-image models has unfolded through various architectural stages, each addressing the specific shortcomings of its predecessors. Starting with the early GAN-based models that established the foundation, to the recent rise of diffusion models, each phase has introduced significant improvements in resolution, quality, and image generation speed. GANs have advanced to generate high-resolution images and are experiencing a resurgence due to their computational efficiency. Autoregressive models have proven effective in zero-shot learning and handling complex prompts, though they are slower in speed. Non-autoregressive transformer models present promising possibilities for faster parallel image generation. Diffusion models, offering high fidelity and diverse images, have become the dominant approach in the field.

Despite these advancements, a significant limitation remains, i.e. current models struggle to accurately generate novel concepts that are not well-represented in the training data. For instance, generating images of cyanobacterial blooms in lentic water bodies— a complex phenomenon that is underrepresented in existing datasets— highlights this challenge. This gap underscores the necessity of integrating customized fine-tuning and semantic enhancement strategies.

In this study, SDXL (Podell et al., 2023) has been chosen as the base model, since it offers an output resolution of 1024×1024 pixels, is open source, and allows both fine-tuning and model customization, essential features to address the specific problem of imaging rare or poorly represented environmental phenomena.

2.2. Personalized image synthesis

As discussed in the previous section, the field of text-to-image synthesis has experienced remarkable progress, driven by advances in deep generative models. Architectures ranging from GANs, through autoregressive models, to more recent diffusion models, have demonstrated an increasing ability to generate high-fidelity, high-diversity images from textual descriptions. These large-scale models, trained on extensive data sets, can synthesize a wide range of concepts, styles, and scenes with remarkable visual quality.

However, despite their impressive generative capabilities, standard text-to-image models face limitations when tasked with generating specific, customized concepts that were not included in the data used for large-scale training (Ruiz et al., 2023). In many application contexts, including the present study, which aims to generate images of unique concepts like cyanobacterial blooms in lentic aquatic environments based on textual prompts, standard models fall short. These models often struggle to accurately represent such highly specialized concepts, as the information provided in a textual description alone is inadequate to achieve the required level of visual fidelity.

To overcome this limitation, model customization techniques have been developed. These techniques involve adapting or extending a pre-trained text-to-image model to integrate and generate specific user-defined concepts, typically using a small set of reference images. The primary goal of these approaches is to enable the model to accurately represent the target concept in new contexts defined by textual prompts, thereby expanding the model's vocabulary with user-specific content (Kumari et al., 2023).

Table 2 provides a comparative summary of the most prominent personalization techniques developed for text-to-image models. Each technique included is analyzed along several critical dimensions: the name of the technique and corresponding literature reference, the base generative architecture on which it is implemented (mostly diffusion models), the type of personalization it allows (single, multiple, pairwise concept), the core technical mechanism it employs, its main contribution to the field, the typical applications it targets, and the inherent limitations identified in the literature.

These customization techniques for text-to-image models are categorized according to the nature of the customization task. They range from single concept customization, such as Textual Inversion (Gal et al., 2022) and DreamBooth (Ruiz et al., 2023), to more complex multi-concept customization such as Custom Diffusion (Kumari et al., 2023), Concept Fusion (Tran et al., 2025), Concept Conductor (Yao et al., 2024)

Table 2
Comparative analysis of personalization techniques for generative text-to-image models.

Technique Name	Organization	Base Architecture Customized	Type of Customization	Main Contribution	Primary Applications	Key Limitations
Textual Inversion (Gal et al., 2022)	Tel Aviv University, NVIDIA	Latent Diffusion Models (e.g., Stable Diffusion)	Single-Concept	Represents new concepts with new "words" in the embedding space of a frozen text-to-image model.	Personalizing text-to-image generation for specific concepts (e.g., objects, styles)	May require multiple images for effective learning; potential for overfitting or underfitting.
DreamBooth (Ruiz et al., 2023)	Google, Boston University	Fine-tuning of pretrained diffusion models (e.g., Stable Diffusion)	Single-Concept	Synthesize novel images of a subject in diverse contexts with high fidelity using a unique identifier.	Subject-driven generation, recontextualization, artistic rendering.	Sensitive to training hyperparameters; can overfit easily; requires significant computational resources.
Custom Diffusion (Kumari et al., 2023)	Adobe Research, Carnegie Mellon University	Text-to-image diffusion model (e. g., Stable Diffusion)	Single & Muti-Concept	Efficiently customizes models for multiple concepts and enables their composition in novel settings.	Synthesizing images with multiple user-specific concepts (e.g., family, pets)	May struggle with composing dissimilar concepts or handling more than two concepts simultaneously.
Concept Fusion (Tran et al., 2025)	University of Information Technology, University of Science	Text-to-image diffusion model	Multi-Concept	Reduces overfitting & attribute leakage for similar concepts via compositional data augmentation.	Multi-concept gen., esp. for similar subjects.	Complexity of separation/ recombination; May not fully solve leakage; Effectiveness depends on augmentation quality.
Concept Conductor (Z. Yao et al., 2024)	Beijing University of Posts and Telecommunications	Text-to-image diffusion model	Multi-Concept	Prevents attribute leakage via isolation; Corrects layout via guidance; Injects concepts via masks/ fusion.	Accurate multi-concept composition with high fidelity & layout control, even for many/similar concepts.	Requires pre-trained single-concept models; High implementation complexity; Performance depends on components.
MC2 (Jiang et al., 2024)	Harbin Institute of Technology	Text-to-image diffusion model (Heterogeneous)	Multi-Concept	Flexible multi-concept composition at inference; Supports heterogeneous models; No joint training needed.	Dynamic composition of multiple concepts at inference time.	Slower inference due to optimization; Effectiveness depends on optimization & model compatibility; Potential conflicts.
MC-LLaVA (An et al., 2024)	Peking University, Intel Labs	Vision-Language Models (VLM)	Multi-Concept (for VLMs)	First multi-concept personalization for VLMs; Efficient joint training via visual init.; Dataset provided.	Enhancing VLMs for multi-concept understanding & response generation (VQA, captioning).	VLM-specific (text-to-image applicability needs study); Depends on dataset quality; Potential VLM forgetting issues.
Pair Customization (Jones et al., 2024)	Carnegie Mellon University, Northeastern University	Stable Diffusion XL with LoRA for style and content	Pairwise (style transfer)	Learns stylistic differences from a single image pair for style transfer while preserving content.	Applying learned styles to new images while maintaining structure.	Struggles with categories significantly different from training; relies on test-time optimization; may fail to maintain input structure.

and MC2 (Jiang et al., 2024). There are also specialized approaches, such as Pair Customization (Jones et al., 2024), aimed at learning stylistic differences between image pairs, which is useful for style transfer tasks and content preservation. Despite their differing methodologies, all of these approaches share a common objective: to enable the model to accurately represent specific concepts in contexts derived from textual descriptions.

Textual Inversion allows customization of the model by learning new word embeddings for user-supplied concepts using 3 to 5 images. This technique introduces new pseudo-words into the model's embedding space, enabling natural language composition for custom creation.

Meanwhile, DreamBooth, is based on the fine-tuning of pre-trained text-to-image diffusion models, establishing a match between a unique identifier and a specific subject, also from a few input images (typically 3–5). The fine-tuning process involves adjusting the weights of the diffusion model by incorporating a specialized pre-preservation loss function, which guarantees high fidelity in concept representation. This technique enables the synthesis of photorealistic images of the new concept across various scenes, poses, viewpoints, and lighting conditions.

Custom Diffusion focuses on multi-concept customization, allowing the extension of text-to-image models, such as Stable Diffusion, to handle multiple concepts simultaneously. The main innovation of this approach lies in identifying that finely tuning only a small and specific subset of the model parameters is sufficient to learn new concepts effectively.

In this study we chose DreamBooth to perform the fine-tuning of the SDXL text-to-image model because of its ability to generate variations of the novel concept (cyanobacterial blooms), including location changes, pose modifications, structure, etc. Most importantly, it preserves the identity of the input concept, that is, to ensure that the generated images capture and preserve very accurately the visual characteristics of the specific concept of the reference images. This technique requires a limited set of X images of the same concept, all of them conditioned by the same text y_s. The conditional y_s takes a simple format, as shown in Equation (1), where the new concept represented by a unique identifier and a class descriptor of the concept are denoted by [unique identifier] and [class name], respectively. The model is trained to associate the unique identifier with the concept illustrated in X. Details of the incorporation of cyanobacterial blooms into the SDXL model using the DreamBooth method can be found in Section 3.3.

$$y_s = \text{"a photo of } [\textit{unique identifier}]\ [\textit{class name}]\text{"} \tag{1}$$

2.3. Prompt engineering for text-to-image generation

The concurrent advancements in LLMs (Brown et al., 2020; Touvron et al., 2023) and text-to-image diffusion models (Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022) have spurred the

development of Prompt Engineering as a critical discipline for effective model interaction. This field focuses on the systematic design and refinement of input prompts to optimize the performance and accuracy of large-scale generative models. Within the domain of synthetic image generation, sophisticated prompt engineering has demonstrated significant potential for controlling images outcomes (Liu & Chilton, 2022; Ramesh et al., 2021).

However, using LLMs to automatically generate high-quality prompts tailored for text-to-image models presents distinct challenges. A successful text-to-image prompt produced by an LLM must meet several essential criteria. First, it should be highly visually specific, offering detailed descriptions of the main subject, its attributes, actions, the surrounding environment, artistic style, and lighting conditions. Second, it must adhere to a clear structure, often requiring the strategic use of keywords recognized by the target text-to-image model to ensure optimal interpretation. Third, the prompt generation process should encourage creativity, allowing for the exploration of diverse visual outputs. Lastly, the prompt must align with the specific characteristics and syntactic preferences of the target text-to-image model.

Manually crafting prompts that consistently meet these criteria is an iterative and labor-intensive process. As a result, there is a strong need for systematic approaches that utilize LLMs to automate or assist in generating diverse and effective text-to-image prompts. Several established prompt engineering techniques can be adapted for this purpose, each with its own advantages and disadvantages in terms of visual specificity, creativity, structural consistency, and ease of implementation. Table 3 provides a comparative analysis of prominent techniques evaluated against these dimensions relevant to text-to-image prompt generation.

Zero-shot Prompting (Radford et al., 2019) involves giving a direct instruction to the LLM without the need for task-specific. Although straightforward to implement ("Very Easy", Table 3), relying solely on the pre-trained knowledge of LLMs (Sahoo et al., 2025; Vatsal & Dubey, 2024) often leads to prompts that lack the required detail and structure for generating complex images ("Low" Visual Specificity and Adhesion Structure, Table 3).

Few-Shot / In-Context Learning (ICL) (Brown et al., 2020) improves prompt effectiveness by including a few examples ("shots") within the prompt itself (Sahoo et al., 2025). These examples guide the LLM towards the desired output format and level of detail. As indicated in Table 3, this technique greatly enhances Visual Specificity and Adhesion Structure ("High"), making it effective for generating well-structured and detailed prompts. Its Potential Creativity is moderate, influenced by the provided examples, and the complexity of implementation is also "Moderate", requiring careful selection of relevant examples.

Role Prompting (Wang et al., 2024) directs the LLM to take on a specific persona or level of expertise (e.g., "Act as a detailed nature photographer") (Schulhoff et al., 2025). This simple technique ("Very Easy" Implementation) can shape the style and focus of the generated prompt, potentially enhancing thematic consistency and detail ("Moderate-High" Visual Specificity, Table 3), though structural adherence may vary ("Moderate").

Self-Refine (Madaan et al., 2023) utilizes an iterative process in which the LLM generates an initial prompt, critiques it based on given criteria (acting as its own feedback mechanism), and then refines the prompt. This approach holds considerable potential for generating highly detailed and creative prompts ("High" across Visual Specificity, Potential Creativity, and Adhesion Structure, Table 3). However, effectively implementing this multi-step process can be challenging ("Difficult" Implementation).

Chain-of-Thought (CoT) (Wei et al., 2022) encourages the LLM to generate intermediate reasoning steps before producing the final output (Sahoo et al., 2025). While this approach is highly effective for tasks that require logical deduction and can help ensure comprehensive prompt elements ("High" Adhesion Structure, Table 3), its linear reasoning process may offer limited benefits for exploring diverse visual concepts ("Low" Potential Creativity) compared to its complexity ("Moderate" Implementation).

Tree of Thoughts (ToT) (Yao et al., 2023) extends CoT by allowing the LLM to explore multiple reasoning paths or prompt variations concurrently, organized in a tree structure. It evaluates different "thoughts" or steps and pursues the most promising branches. This approach excels in problems requiring exploration of diverse possibilities, offering high creative potential ("High" Potential Creativity, Table 3), but it comes with significant implementation complexity ("Very Difficult").

Considering the goal of this study –to generate a diverse set of synthetic images of cyanobacterial blooms and navigational obstacles in lentic water bodies using SDXL fine-tuned by DreamBooth– selecting the right LLM prompting technique is essential. The prompts generated by LLaMa 2 (Touvron et al., 2023) must be sufficiently detailed (high visual specificity) and well-structured for the text-to-image model, while also enabling the generation of diverse scenarios (potential creativity). Additionally, it is important to strike a balance between effectiveness and implementation feasibility.

Upon evaluating the techniques presented in Table 3, Few-Shot / In-Context Learning (ICL) emerges as the most suitable approach for our requirements. Its strength lies in offering explicit examples ("shots") that guide the LLM (LLaMa 2) to produce outputs with the desired level of detail and structure ("High" Visual Specificity and Adhesion Structure). By carefully curating diverse examples that represent various bloom characteristics and environmental settings, we can direct the LLM to generate a broad set of high-quality prompts. While techniques like Self-Refine and ToT offer greater theoretical creativity, their implementation complexity ("Difficult" to "Very Difficult") makes Few-Shot / ICL ("Moderate" Implementation) a more practical choice for achieving controlled diversity and detail in this specific application. Role Prompting, although easy, offers less consistency in structure, and Zero-shot is insufficient for the required detail. CoT is less suited to the goal of visual diversity compared to ICL. Therefore, we have chosen the Few-Shot / ICL technique to generate diverse and descriptive prompts for synthesizing images of cyanobacterial blooms and navigational obstacles in lentic aquatic environments, as outlined in Section 3.4.

In summary, while previous works have advanced the fields of text-to-image synthesis, model fine-tuning, and prompt engineering individually, our proposal stands out by integrating these techniques within a practical and domain-specific context. Unlike existing studies that

Table 3
Comparative analysis of different techniques for the task of generating prompts oriented to text-to-image models.

Technique	Visual Specificity	Potential Creativity	Adhesion Structure	Ease of Implementation
Zero-shot (Radford et al., 2019)	Low	Low	Low	Ver Easy
Few-Shot/ In-Context Learning (ICL) (Brown et al., 2020)	High	Moderate	High	Moderate
Role Prompting (Wang et al., 2024)	Moderate-High	Moderate	Moderate	Very Easy
Self-Refine (Madaan et al., 2023)	High	High	High	Difficult
Chain-of-Thought (CoT) (Wei et al., 2022)	High	Low	High	Moderate
Tree-of-Thoughts (ToT) (S. Yao et al., 2023)	Moderate	High	Moderate	Very Difficult

focus on general-purpose or artistic applications, aur approach introduces: (i) fine-tuning of the SDXL model using DreamBooth with real cyanobacterial bloom patches (an environmental phenomenon that is rarely represented), (ii) automatic generation of high quality promts through LLaMa 2 to enhance scene diversity. This combined strategy enables the synthesis of realistic and diverse aquatic scenes, specifically designed for the training DL models on ASVs, addressing critical limitations in data availability for real-world monitoring of cyanobacterial blooms and navigation safety. This integration represents the core innovation of our work.

3. Method

We present a novel methodology that leverages the SDXL model, fine-tuned via the DreamBooth technique, to synthesize high-fidelity images depicting complete scenes of lentic water bodies. To guide the image generation process, we employ LLaMA 2 for the construction of semantically rich and diverse natural language prompts. The generated images are subsequently annotated using the Grounding DINO model (Liu et al., 2023), which identifies and localizes scene elements by producing bounding boxes around relevant objects. These bounding boxes are then processed by the Segment Anything Model (SAM) (Kirillov et al., 2023), to generate precise segmentation masks for each identified object. The resulting annotations undergo manual validation and refinement to ensure accuracy. The curated, annotated dataset is used to train a convolutional neural network (CNN)-based dual-task architecture that simultaneously performs object detection and semantic segmentation. This section provides a detailed overview of the datasets utilized, the architecture of the CNN-based dual-task model, the prompt generation strategy, the image synthesis pipeline, and the evaluation metrics adopted.

3.1. Dataset

This study utilizes two distinct datasets. The first dataset, originally introduced in Barrientos-Espillco et al. (2023), comprises cyanobacterial bloom patches extracted from real-world imagery capturing bloom occurrences in lake or reservoir water bodies. These patches undergo a series of data augmentation procedures and are subsequently composited onto background images, as outlined in Barrientos-Espillco et al. (2023). The extraction methodology is also thoroughly described in that study. These patches are used for the fine-tuning of the SDXL text-to-image diffusion model. The detailed fine-tuning procedure is presented in Section 3.3. Representative examples of the cyanobacterial patches from this dataset are illustrated in Fig. 2.

The second dataset, as detailed in Barrientos-Espillco et al. (2024), comprises a total of 3,286 training images and 822 validation images depicting a wide range of aquatic environments, including lakes, lagoons, reservoirs, wetlands, and inland waterways. These images include objects typical of such environments, such as submerged rocks, floating trunks, aquatic vegetation, swimmers, canoes, birds, among others, as well as various textures such as water body, grasses, trees, mountains, sky and, of course. Importantly, cyanobacterial blooms are also present within the water bodies. The dataset is annotated at both the object level (for detection tasks) and the pixel level (for semantic segmentation). In the present study, this dataset is employed for training and evaluating the dual-task model, which concurrently performs object detection and semantic segmentation. Representative examples from this dataset are shown in Fig. 3.

3.2. Dual-task model

To assess the suitability of the synthetic images generated by the customized SDXL text-to-image model, we employ the CNN-based architecture outlined in our previous study (Barrientos-Espillco et al., 2024). This architecture consists of three components: Backbone, Neck, and Head, as depicted in Fig. 4. The Head is further split into two branches, one of which focuses on object detection, following the foundational structure established in YOLOv3, and also utilized in YOLOv4 (Bochkovskiy et al., 2020) and YOLOv5. The selection of YOLOv3 is driven by its speed, ease of use, and accuracy. While more advanced versions of YOLO, such as YOLOv8 and subsequent releases, focus on spatial attention, this aspect is thoroughly addressed by the Neck's Convolutional Block Attention Module (CBAM) (Woo et al., 2018) in the proposed architecture for object detection. However, this attention mechanism is not effective for detecting amorphous patches, where it has little relevance. In addition, YOLOv3 is less computationally intensive than newer versions such as YOLOv8. This clearly favors its use in systems with limited hardware resources, including memory usage, such as MSV on board an ASV. On the other hand, several experiments have shown that the performance in terms of success during the inference process does not show significant improvements compared to YOLOv3, as reported in section 4.1. This justifies the choice of YOLOv3 as an appropriate option in the proposed approach.

The second branch is dedicated to semantic segmentation, incorporating three semantic networks —BiSeNet (Yu et al., 2018),

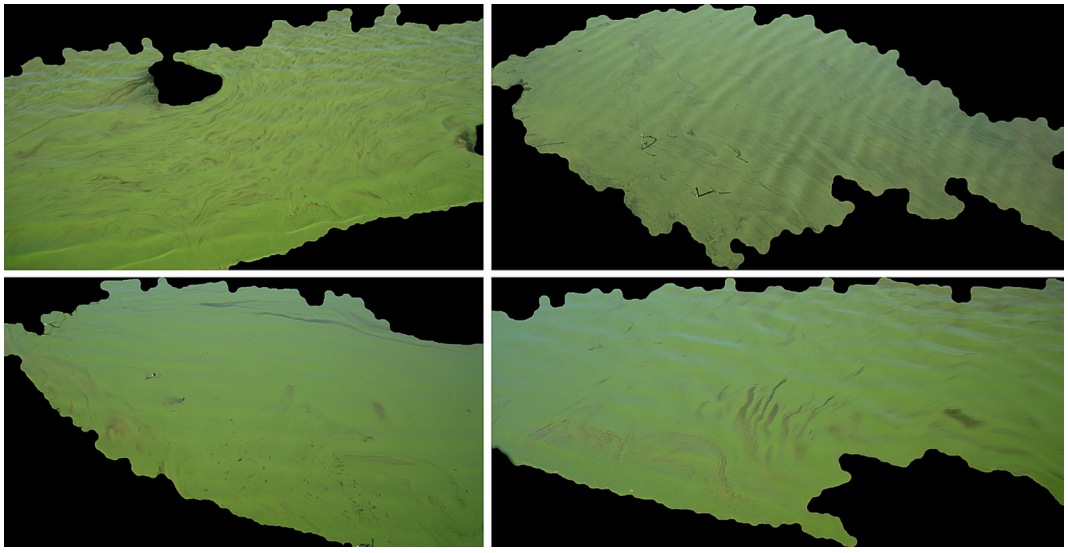

Fig. 2. Sample of real patches of cyanobacterial blooms.

Fig. 3. Some images of the dataset presented in the study by Barrientos-Espillco et al. (2024).

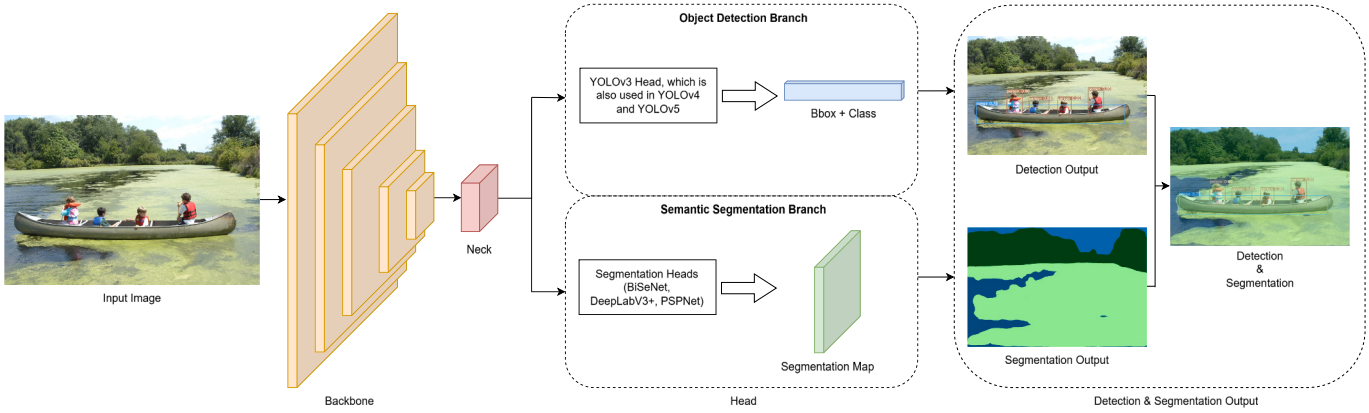

Fig. 4. Overview of the architecture of the dual-task model integrating object detection and semantic segmentation (Barrientos-Espillco et al., 2024).

DeepLabV3+ (L.-C. Chen et al., 2018), and PSPNet (Zhao et al., 2017)—for performance comparison. In this study, only the PSPNet network (head) was implemented in the semantic segmentation branch, as it demonstrated the best performance according to the result reported in Barrientos-Espillco et al. (2024). Moreover, the architecture incorporates the CBAM to direct attention to challenging areas in adverse environmental conditions.

3.3. Stable diffusion XL fine-tuning

The goal of this proposal, as previously mentioned, is to generate synthetic images of lentic waters that feature navigational obstacles, cyanobacterial blooms, and the surrounding environment, such as grasses, trees, mountains, among others. Due to the limitation of the SDXL model in generating blooms within the water body, we fine-tuned it using the DreamBooth technique. This method efficiently incorporates a custom concept into the model, enabling the generation of a wide range of images of the specific concept in various contexts, and ensuring its seamless integration into the scene with just a few reference images. In our study, 10 real cyanobacterial patches were carefully selected from the first dataset (see Section 3.1) to fine-tune the SDXL model, enabling the generation of new, accurate images.

To guarantee the accurate representation of the new concept, cyanobacterial blooms, a distinct and uncommon identifier was assigned to prevent any overlap with concepts typically found in the SDXL model. The identifier chosen was “hcb”, and the class name was set as “cyanobacterial blooms”. Using this, we constructed our instance prompt, which resulted in the phrase: *a photo of [unique identifier] [class name] … (“a photo of hcb cyanobacterial blooms …”)*.

3.4. Prompts generation

A systematic process was implemented to create descriptive and diverse input prompts for generating a set of synthetic images depicting cyanobacterial blooms and navigation obstacles in lentic aquatic environments using the SDXL text-to-image diffusion model, fine-tuned with

DreamBooth. The objective was to ensure wide variability in the generated images in terms of composition, obstacle types, bloom characteristics, and environmental conditions.

LLaMa 2 was employed as the primary tool for the automated generation of prompts. Based on the comparative analysis of prompt engineering techniques presented in Section 2.3, the Few-Shot / In-Context Learning (ICL) technique (Brown et al., 2020) was adopted. This choice is based on ICL's ability to effectively guide the LLM towards generating text that adheres to specific formats and desired levels of detail, by providing concrete examples ("shots") within the context of the instruction prompt to the LLM.

Two exemplary "shots" were meticulously designed to instruct LLaMa 2, encapsulating the desired structure, information density, and keywords for the final prompts. These guiding "shots" were as follows:

Shot 1: Obstacle Example (No Cyanobacterial Bloom)

Low angle water level view from a canoe gliding slowly near the shore of a calm, pristine mountain lake. Several large, jagged granite boulders rise partially from the crystal-clear water, their submerged parts clearly visible below the surface. Bright, direct sunlight creates strong highlights and deep shadows, emphasizing the wet texture of the ***rocks***. ***Sharp focus*** captures every crack and crevice on the nearest boulder. Subtle ripples reflect the clear blue sky. Distant pine trees line the shore. ***Hyper-realistic***, ***highly detailed*** composition. Award-winning nature photography, Flickr style. Shot on Canon EOS R5.

Shot 2: Obstacle with Cyanobacterial Bloom Example

Eye-level perspective from an Autonomous Surface Vehicle (ASV) navigating through a dense [***hcb***] [***cyanobacterial blooms***] covering the surface of a eutrophic pond. A large, weathered log floats centrally, partially entangled with thick aquatic weeds (like milfoil). The water is completely obscured by a thick, paint-like layer of vibrant blue-green algae scum, forming swirling patterns around the log. ***Sharp focus*** is maintained on the ***highly detailed***, peeling bark of the ***trunk***, contrasting dramatically with the smooth texture of the bloom. Diffused overcast lighting creates soft, even illumination with minimal reflections on the algal surface. Background shows dense reeds at the pond's edge. ***Hyper-realistic*** representation of a severe algal bloom. Environmental photography, Flickr style. Captured by Sony A7R III.

These "shots" explicitly incorporated: i) a detailed description of the primary subject (obstacle) and its attributes; ii) specific characteristics of the lentic aquatic environment (type of water body, surface state, initial water clarity, surrounding vegetation); iii) precise indication of the viewpoint, simulating a low-angle or water-level shot from a canoe or an ASV; iv) lighting and atmospheric conditions (time of day, weather); v) keywords to reinforce visual style and quality: hyper-realistic, highly detailed, sharp focus; vi) photographic style guidelines associated with high-quality platforms (Flickr style photo, award-winning photography); vii) simulated camera metadata (e.g., Shot on Canon EOS R5); and viii) in the case of the second "shot", a detailed description of the cyanobacterial blooms (density, texture, color, pattern).

The generation process involved providing these two "shots" as context to Llama 2, as outlined in Table 4, followed by instructions to generate new prompts that followed the same structural and semantic template. The creation of variations was requested for a predefined list of 7 classes of common obstacles in lentic environments (partially submerged rocks, floating tree trunks, dense aquatic vegetation, swimmers, canoes, paddles, birds) and various scenarios of cyanobacterial bloom presence and density (ranging from clear water to dense and extensive surface layers). Table 5 presents a selection of prompts generated by LLaMa 2 based on the structural and semantic template.

This methodological approach based on Few-Shot/ICL enabled the generation of a diverse and controlled corpus totaling 140 prompts. Specifically, 70 prompts were generated depicting the 7 obstacle types in lentic aquatic environments without visible blooms (10 prompts per obstacle class), and an additional 70 prompts incorporated variable descriptions of cyanobacterial blooms alongside each obstacle type (10 prompts per combined class). This diverse set of prompts was specifically designed to steer the text-to-image model towards the synthesis of the varied and realistic images required for the study's objectives.

Table 4
Summary of instruction provided to LLaMa 2 for the generation of text-to-image prompts.

Instruction component	Detail / Specification provided to LLaMa 2
Role and objective	• Act as an expert prompts engineer for text-to-image (SDXL) models. • Objective: Generate diverse and detailed prompts to create synthetic training data (navigation obstacle detection in lentic aquatic environments).
1. Technique and examples guide	• Specified technique: Few-Shot / In-Context Learning (ICL). • 2 detailed "shots" (examples) were provided as a strict guide: - Shot 1: Obstacle (rocks), clear lentic environment, no bloom. - Shot 2: Obstacle (trunk/vegetation), lentic environment with dense cyanobacterial bloom.
Generation task 1	• Title: Prompts without cyanobacterial blooms. • Instruction: Generate 10 unique prompts for each of 7 kinds of obstacles (Rocks, Trunks, Vegetation, Swimmers, Canoes, Paddles, Birds) in clear lentic environments. • Total: 70 prompts. • Guidance: Follow structure/detail of Shot 1.
Generation task 2	• Title: Prompts with cyanobacterial blooms. • Instruction: Generate 10 unique prompts for each of the 7 obstacle classes. Incorporate varying degrees of cyanobacterial blooms (detailed descriptions), realistically integrated into the environment. • Total: 70 prompts. • Guidance: Follow structure/detail of Shot 2.
Mandatory restrictions (for all 140 prompts)	• Strictly apply the following requirements demonstrated in the shots: - Point of view: Low angle/eye level from canoe/ASV (Avoid aerial views). - Environment: Detailed description of lentic environment (water, shore, weather, light). - Realism keywords: Include hyper-realistic, highly detailed, sharp focus. - Style keywords: Include Flickr style or similar (e.g. award-winning photography). - Camera Model: Include random camera tag from predefined list (Canon EOS R5/R6, Sony A9/A6400/A7R III, Nikon D850, Fujifilm X-T4). - Details: Obstacle specific description; ensure overall diversity in each category. - Structure: Follow descriptive pattern of guide shots.
Required output format	• Deliver the prompts directly as a list. • Group first the 70 of Task 1 (by obstacle type). • Group after the 70 of Task 2 (by obstacle type).

3.5. Image generation

In this study, we generated synthetic images of navigational obstacles and cyanobacterial blooms in lentic water bodies using the custom text-to-image model of SDXL. To achieve high-quality and diverse synthetic images, we used prompts specifically designed for this model, as outlined in Section 3.4. For each type of navigation obstacle, we generated 100 synthetic images, with 10 images per prompt. Fig. 5 displays some of these synthetic images for each navigational obstacle. In total, we produced 700 synthetic images for the different obstacles. Likewise, 700 synthetic images depicting cyanobacterial blooms were generated, as illustrated in Fig. 6.

To train the dual-task model, we manually selected the best synthetic images. Out of the 700 images of navigational obstacles, 300 were chosen, and similarly, 300 images were selected from the 700 images of cyanobacterial blooms. As a result, the final training set consisted of 600 synthetic images generated using the SDXL text-to-image model.

It is important to note that, while the number of images used may seem relatively small from a Deep Learning perspective, it is more than

Table 5
Specific examples of prompts generated by Llama 2 following the guidance of the "shots".

Class	Prompt
Based on Shot 1 (Obstacles without cyanobacterial bloom)	
Rock	*"Low angle perspective from an ASV moving slowly near large, smooth sedimentary **rocks** partially submerged at the edge of a calm reservoir under a bright, cloudless sky. Crystal-clear water allows visibility of the rocky bottom near the shore. **Sharp focus** on the **highly detailed** patterns and wet surfaces of the rocks. Reflections are crisp on the still water. **Hyper-realistic**, highly detailed geological feature photo. Landscape photography, Flickr style. Shot on Nikon D850."*
	*"Water level view from a stationary canoe very close to moss-covered **rocks** emerging from a clear, tannin-stained pond. Overcast, diffused lighting creates soft highlights on the wet moss. **Sharp focus** captures the intricate texture of the moss and the dark, still water. Dense forest visible on the far bank. **Hyper-realistic**, **highly detailed** composition. Quiet nature scene, Flickr style. Shot on Canon EOS R6."*
Trunk	*"Low angle water level view from a kayak showing a massive, ancient waterlogged **trunk** floating near the reedy edge of a calm, mist-covered lake at sunrise. The glassy water surface perfectly mirrors the log and the soft pastel colors of the dawn sky. **Sharp focus** captures the **highly detailed**, gnarled texture of the dark, wet wood and individual strands of mist swirling above the water. **Hyper-realistic**, highly detailed atmospheric scene. Serene landscape photography, Flickr style. Shot on Fujifilm X-T4."*
	*"Eye-level perspective from inside a canoe looking towards a cluster of smaller, sun-bleached **branches** floating in the clear, shallow water of a lagoon. Bright midday sun creates dazzling reflections on the water surface. **Sharp focus** on the highly detailed texture of the bleached wood against the blue water. Sandy bottom visible. **Hyper-realistic**, **highly detailed** summer scene. Coastal life photography, Flickr style. Shot on Sony A6400."*
Based on Shot 2 (Obstacles with cyanobacterial bloom):	
Rock + Blooms	*"Eye-level view from an ASV navigating a narrow channel in a lake where smooth, dark basalt **rocks** emerge from water covered by a moderately thick layer of of swirling green and blue-green **[hcb] [cyanobacterial blooms]**. The bloom parts slightly around the rocks, revealing murky water beneath. Overcast sky provides soft, even lighting, emphasizing the slimy texture of the bloom. Sharp focus highlights the contrast between the **highly detailed** wet rock surface and the adjacent algal mat. **Hyper-realistic** depiction of bloom encroachment. Environmental documentation, Flickr style. Captured by Nikon D850."*
Trunk + Blooms	*"Low angle close-up from a kayak beside a large, partially submerged decaying tree **trunk** in a stagnant pond area. The water surface is almost entirely covered by a dense, lumpy crust of dark green **[hcb] [cyanobacterial blooms]**, resembling thick sludge, clinging heavily to the visible parts of the trunk. Direct, harsh midday sun reflects minimally off the bloom. Sharp focus captures the **highly detailed**, rotten wood texture peeking through the oppressive algal layer. **Hyper-realistic** scene of advanced eutrophication. Impactful environmental photo, Flickr style. Shot on Canon EOS R5."*

sufficient to validate the proposed strategy based on the report results. This demonstrates the strategy's essential role in generating synthetic images for the given context and, more broadly, for other potential contexts, as will be discussed in the final conclusions.

3.6. Implementation details

Image generation. We utilize SDXL as a pre-trained text-to-image model, widely used due to its open-source availability, and fine-tuning it using DreamBooth. We use the third-party implementation of huggingface (von Platen et al., 2024) for DreamBooth. The training is conducted with a batch size of 8, a learning rate of 5x10^-5, and a total of 1000 training steps.

Dual-task model. We adapted the configuration implemented in the study by Barrientos-Espillco et al. (2024). Specifically, we applied the Stochastic Gradient Descent (SGD) optimizer (Bottou, 2010), with a learning rate of 5e-2, a decay of 5e-4, and a momentum of 0.90. These hyperparameters were previously determined through a grid search procedure, where multiple configurations were evaluated to optimize the convergence and performance of the model in the same application context. Given the proven stability and effectiveness of this configuration in that study, we applied it directly in the present work. The training process lasted for approximately 100 epochs on the object detection and semantic segmentation datasets. Table 6 summarizes the main hyperparameters used to train the dual-task CNN model described in this study.

To further support the suitability of the selected hyperparameters, we include a convergence plot showing the training and validation loss function over 120 epochs. As illustrated in Fig. 7, the model demonstrates a stable decrease in both training and validation loss, with convergence observed around epoch 100. This behavior confirms that the learning rate and other hyperparameters previously tuned in Barrientos-Espillco et al. (2024) remain effective under the current experimental conditions.

3.7. Evaluation metrics

Image generation. We assess the synthetic images generated by SDXL using the dual-task CNN model, which concurrently tackles both the object detection and semantic segmentation tasks.

Dual-task model. We use mean average precision (mAP) with a confidence threshold of 0.5 to evaluate the object detection task, and mean intersection over union (mIoU) as the metric for semantic segmentation.

4. Results and discussion

In this section, we present the results of benchmarking the performance of the dual-task CNN model, which was trained independently on three different datasets. Table 7 summarizes the main characteristics of these datasets. The first dataset comprises 600 randomly selected images from the study Barrientos-Espillco et al. (2024). The second dataset includes 600 synthetic images generated using the SDXL model, fine-tuned with the DreamBooth method. The third dataset combines both data sources (600 real and 600 synthetic images). Performance evaluation was conducted on a single test set of 200 images, randomly selected from the validation set provided in the study Barrientos-Espillco et al. (2024). This experimental approach enables a detailed comparative analysis of how the different training data sources influence the performance of the dual-task model.

Consistent with the study by Barrientos-Espillco et al. (2024), the same backbone architecture (CSPDarknet-53) and object detection method (the YOLOv3 head, as used in YOLOv4 and YOLOv5) were applied in each experiment. For the semantic segmentation branch, the PSPNet semantic network head was implemented, as it demonstrated optimal performance in the results reported in the study Barrientos-Espillco et al. (2024).

Regarding model complexity, the number of convolutional units in the dual-task CNN architecture was selected based on our prior study (Barrientos-Espillco et al., 2024), in which the architecture was optimized for lentic water scenes. This configuration is designed to balance model expressiveness and generalization. During training, no signs of overfitting (e.g., increasing validation loss) or underfitting (e.g., high training loss plateau) were observed. The convergence curves (Fig. 7) show a steady decrease in both training and validation loss functions, reinforcing the suitability of the architecture and its robustness to the dataset size used in this study.

4.1. Quantitative comparison

Table 8 presents the quantitative results of the dual-task CNN model's performance, trained independently on three different image datasets featuring lentic water scenes and evaluated on a single test set. The table highlights (a) the classes of interest for both object detection and semantic segmentation, and (b) columns 2 to 6 show the corresponding performance results. Regarding the object detection branch,

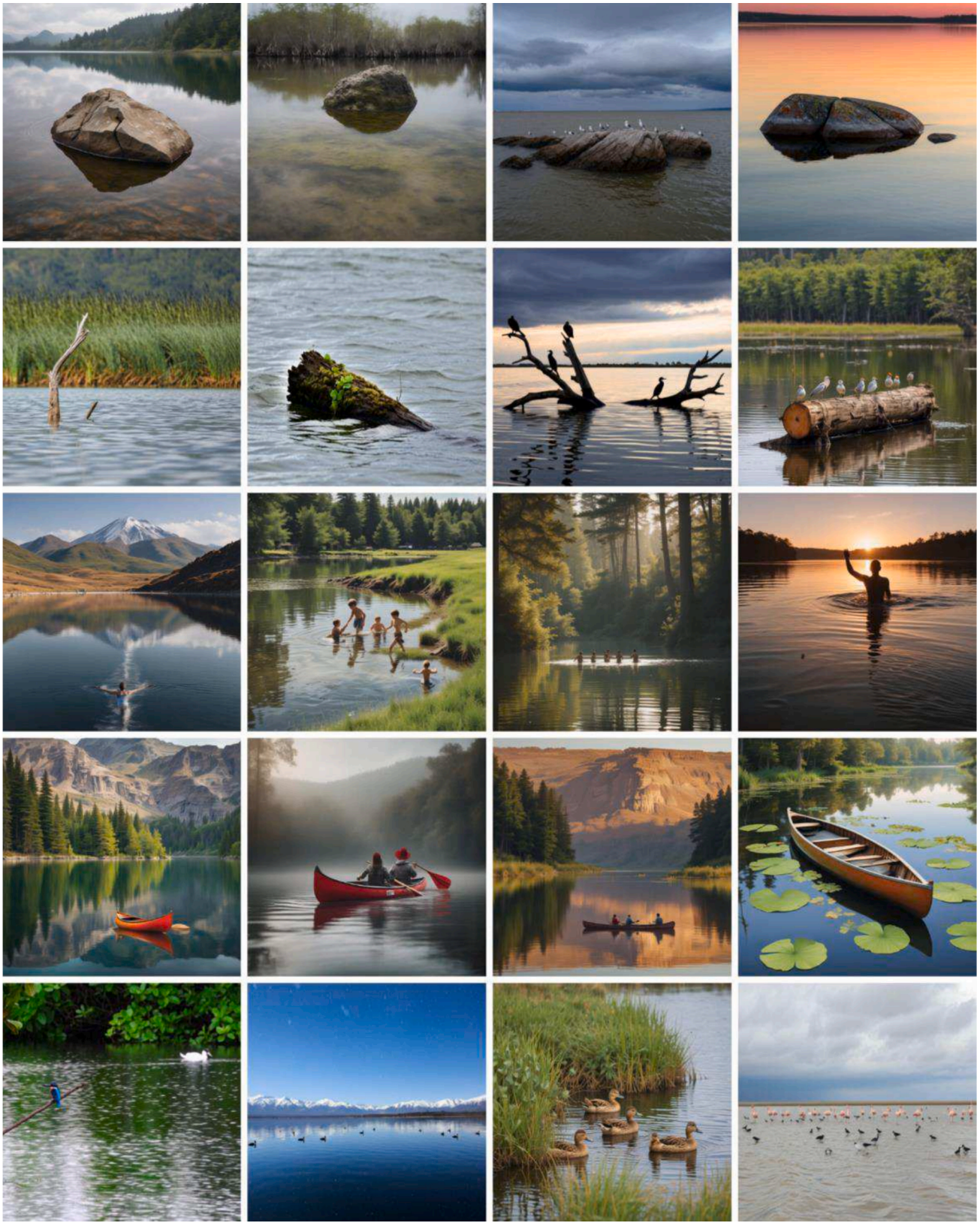

Fig. 5. Synthetic images of obstacles to navigation of Autonomous Surface Vehicles (ASVs) in lentic water bodies generated by the custom Stable Diffusion XL text-to-image model. In the first row, submerged rocks. In the second row, floating trunks. In the third row, swimmers. In the fourth row, canoes and navigators. In the fifth row, birds.

Fig. 6. Synthetic images of cyanobacterial blooms in lentic water bodies generated by the custom Stable Diffusion XL text-to-image model.

Table 6
Summary of training hyperparameters used in the dual-task CNN model.

Hyperparameter	Value
Optimizer	SGD
Learning rate	5e-2
Decay	5e-4
Momentum	0.90
Batch size	16
Number of epochs	100
Input image size	512x512
Loss functions	YOLOv3 loss (object detection), Cross-entropy (segmentation)
Learning rate schedule	$\alpha = \frac{\alpha_0}{1+d+e}, \alpha_0 = 1.5e-3, d = 5e-4, e = 100$

it's important to note that switching from YOLOv3 to YOLOv8 yields virtually no improvement in detection performance—only marginal gains of a few hundredths in accuracy for certain categories (like Canoe and Trunk), with similar minimal changes observed across the remaining categories.

The key highlights of the results presented in Table 8, specifically for the object detection branch, are as follows:

- The dual-task CNN model trained on the Barrientos-Espillco et al. (2024) dataset performs notably well on three classes; "canoe", "person", and "paddle". However, its performance on the classes "dog", "bird", "rock", and "trunk" is not sufficient. This underperformance in certain classes is believed to stem from two factors highlighted in the study Barrientos-Espillco et al. (2024): a) the limited diversity of these classes in the training set, and b) the class imbalance, where the "dog", "bird", "rock", and "trunk" classes are

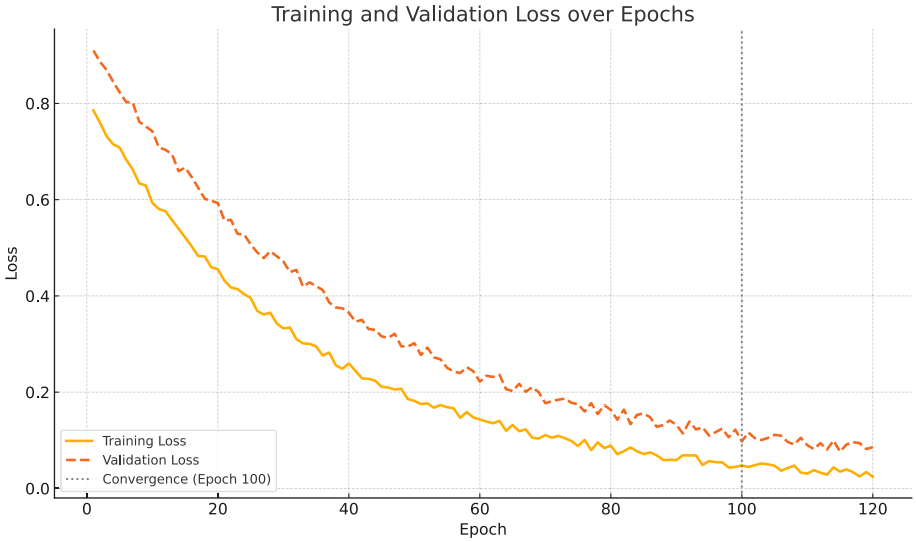

Fig. 7. Training and validation loss over 120 epochs using the selected hyperparameters. The model shows convergence around epoch 100.

Table 7
Summary of the training datasets used in the benchmarking experiments.

Dataset	Type	No. of Images	Classes Included	Image Resolution	Description / Source
Barrientos-Espillco et al. (2024)	Real	600	Cyanobacterial blooms, obstacles, background	512x512 pixels	Random selection from Barrientos-Espillco et al. (2024)
SDXL-DreamBooth	Synthetic	600	Cyanobacterial blooms, obstacles, background	512x512 pixels	Generated using SDXL fine-tuned with DreamBooth + LLM prompts
Hybrid Dataset (Barrientos-Espillco eta al. (2024) + SDXL-DreamBooth)	Real + Synthetic	1,200	Cyanobacterial blooms, obstacles, background	512x512 pixels	Combination of the two datasets above

Table 8
Performance benchmarking results of the dual-task CNN model trained on three different datasets and evaluated on a single test set. The head for object detection is YOLOv3, used in YOLOv4 and YOLOv5 and the head for semantic segmentation is PSPNet.

Datasets Classes	Barrientos-Espillco et al. (2024)		SDXL-DreamBooth		Barrientos-Espillco et al. (2024) + (SDXL-DreamBooth)	
	mAP@.5	mIoU	mAP@.5	mIoU	mAP@.5	mIoU
Object detection branch						
Canoe	87.42	—	87.39	—	88.70	—
Person	96.14	—	96.10	—	97.32	—
Paddle	72.10	—	71.98	—	72.94	—
Dog	32.56	—	54.52	—	54.71	—
Bird	28.48	—	51.10	—	51.33	—
Rock	34.31	—	67.22	—	67.86	—
Trunk	33.22	—	66.11	—	66.91	—
All	54.89	—	70.63	—	71.40	—
Semantic segmentation branch						
Water	—	93.72	—	93.02	—	94.30
Blooms	—	90.10	—	92.80	—	92.84
Grass	—	54.12	—	64.48	—	64.92
Dirt	—	39.93	—	55.82	—	55.95
Tree	—	84.85	—	85.10	—	85.80
House	—	46.91	—	58.74	—	58.86
Mountain	—	43.11	—	54.92	—	54.78
Sky	—	93.14	—	92.90	—	94.82
All	—	68.24	—	74.72	—	75.28

underrepresented compared to others, reflecting their minority status in the training data.

- In contrast, the dual-task CNN model trained on synthetic images generated using customized SDXL outperforms the model trained in Barrientos-Espillco et al. (2024) dataset. This improvement is attributed to the model's ability to effectively address the issues of diversity and class imbalance. When examining mAP per class, there is a notable enhancement compared to the results from the model on the Barrientos-Espillco et al. (2024) dataset. Specifically, the "dog" class shows a remarkable improvement of 21.96 %, followed by the "bird" class with a 22.62 % increase, the "rock" class with a 32.91 % improvement, and the "trunk" class with a 32.89 % gain.
- Furthermore, the dual-task CNN model trained with a combination of the Barrientos-Espillco et al. (2024) dataset and the generated synthetic images delivers exceptional performance, surpassing the results of the models trained on each dataset individually. The classes that benefit most from this dataset combination are "canoe", "person" and "paddle", while the "dog", "bird", "rock", and "trunk" classes show limited or no improvement. The model's overall performance is attributed to the larger and more diverse set of images it was trained on, resulting from the combination of the two datasets.

On the other hand, the analysis of the results for the semantic segmentation branch, as presented in Table 8, can be summarized as follows:

- First, it is important to highlight the satisfactory performance of the model trained on the Barrientos-Espillco et al. (2024) dataset for the semantic classes "water", "sky", "blooms" and "tree". However, the classes with lower performance, such as "grass", "house", "mountain" and "dirt", are precisely those with less representation in the training set, as noted in the study Barrientos-Espillco et al. (2024).
- Second, the model trained on synthetic images generated by SDXL outperforms the model trained on the Barrientos-Espillco et al. (2024) dataset in most semantic classes, particularly those with lower representation, as mentioned earlier. There is a remarkable 15.89 % improvement in the "dirt" class, followed by an 11.83 % increase in the "house" class, an 11.81 % increase in the "mountain" class, a 10.36 % increase in the "grass" class, and a 2.70 % increase in the "blooms" class.
- Third, the model trained with the combination of the Barrientos-Espillco et al. (2024) dataset and the generated synthetic images shows slightly better overall performance than the model trained solely on synthetic images. This improvement is attributed to the larger and more diverse set of images used for training, resulting from the combination of the two datasets.

4.2. Qualitative comparison

Fig. 8 displays the visual results produced by the dual-task CNN model, trained independently on three different datasets. The figure presents the results on a sample image, with the findings being representative and applicable to the entire set of images from analyzed datasets. The first row shows the results from the model trained on the dataset presented in the study by Barrientos-Espillco et al. (2024). In this configuration, the object detection branch performs well in predicting the objects of interest, accurately identifying both bounding boxes and class labels. Simultaneously, the semantic segmentation branch predicts the amorphous structures with a reasonable degree of accuracy.

The second row presents the results from the model trained with synthetic images generated using custom SDXL. In this case, the object detection branch makes more accurate predictions overall, outperforming the model trained on the Barrientos-Espillco et al. (2024) dataset. Similarly, the semantic segmentation branch demonstrates improved prediction accuracy.

The third row shows the results from the model trained using a combination of both datasets (Barrientos-Espillco et al. (2024) and customized SDXL). In this setup, the object detection branch shows a slight improvement over the model trained solely with synthetic images generated by custom SDXL. Likewise, the semantic segmentation branch produces more accurate predictions in general, with sharper boundaries in amorphous segments.

In conclusion, the dual-task CNN model, when trained independently on the three datasets, exhibits no significant differences from a visual and qualitative perspective.

5. Conclusions

This paper presents a novel approach that combines DreamBooth-based fine-tuning of the SDXL model with LLM-driven prompts generation (LLaMa 2) to synthesize realistic images of cyanobacterial blooms and navigational obstacles in lentic water bodies. The generated data addresses the challenge of limited training and validation samples for

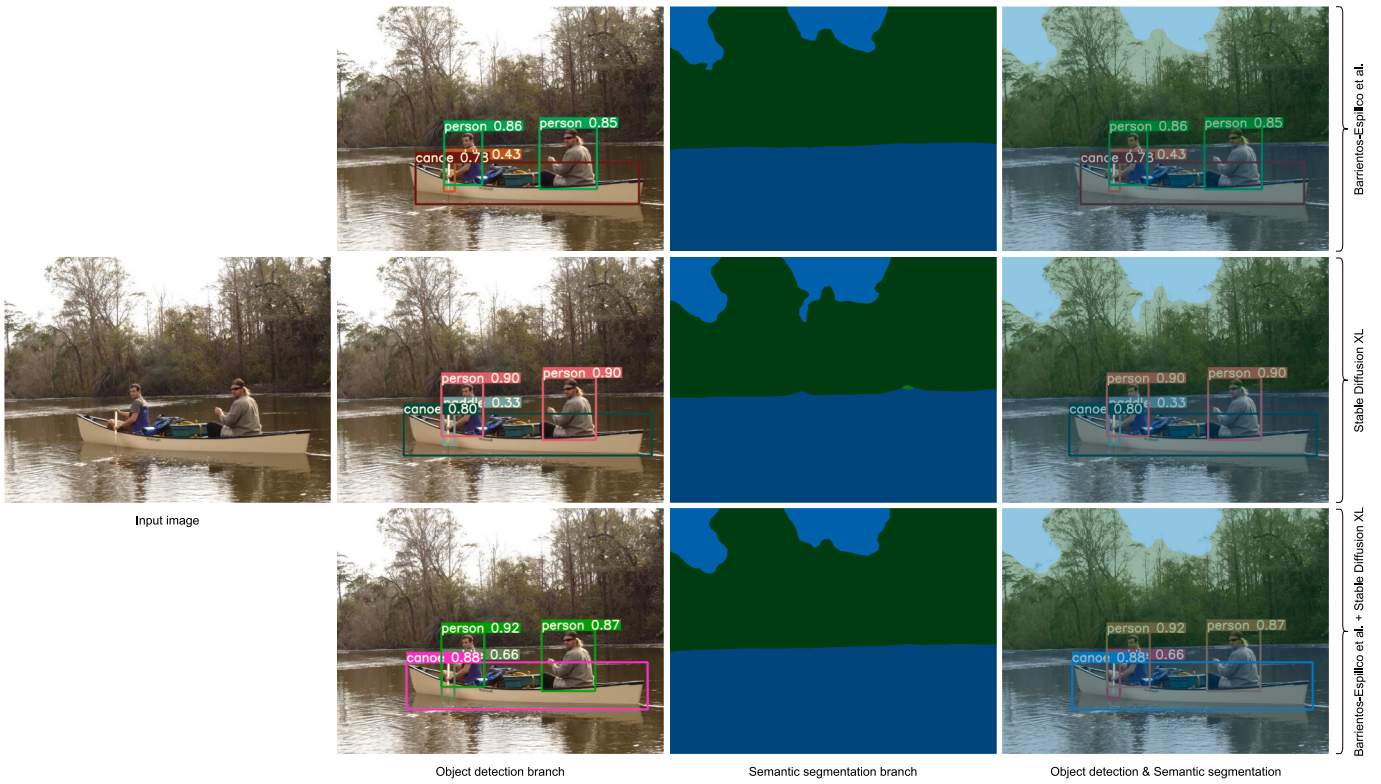

Fig. 8. Visual presentation of experimental results obtained using the dual-task CNN model, trained independently on three different datasets: Barrientos-Espillco et al. (2024), custom Stable Diffusion XL, and a combination of both. In the first row, the results of the model trained exclusively on the Barrientos-Espillco et al. (2024) dataset are shown. The second row presents the results of the model trained with synthetic images generated by Stable Diffusion XL. In the third row, the results of the model trained using the combination of both datasets are illustrated.

convolutional neural networks (CNNs), enhancing their ability to detect harmful blooms and recognize obstacles critical for the ASVs equipped with an MVS system.

Experimental results demonstrate a notable improvement in the performance of minority classes within the real image dataset when realistic synthetic images were incorporated into the training and validation datasets. Specifically, experimental results demonstrate improvements of up to 15.74 % in object detection and 6.48 % in semantic segmentation when using the synthetic dataset, with further gains observed when combining synthetic and real data. The methodology not only enhances model performance in environmental monitoring tasks but also shows strong potential for transferability to other domains such as medical imaging, remote sensing, oil spill detection, and precision agriculture. These results validate the proposed methodology as a practical and scalable solution for generating high-quality training data in domains where collecting real samples is difficult or expensive.

While the results are promising, several key limitations must be considered. First, although the synthetic images generated via SDXL fine-tuned with DreamBooth appear realistic, they may exhibit subtle artifacts or lack certain statistical characteristics inherent in real-world data, potentially limiting the generalization of deep learning models trained exclusively on synthetic samples. Second, the utility of these images is highly dependent on the quality and diversity of the prompts, which are influenced by the LLM's ability to produce semantically rich and varied descriptions. Lastly, our findings indicate that integrating synthetic images with real data yields better performance, highlighting that synthetic data is most effective as a supplement rather than a substitute for real-world datasets.

Future work will explore several promising directions to further advance our approach. First, we aim to improve prompt diversity and semantic control by incorporating more sophisticated prompt engineering strategies, including context-aware methods and iterative refinement techniques. Second, we plan to investigate the application of multimodal diffusion models that integrate textual, spatial, and temporal inputs to generate scene-consistent image sequences, thereby enhancing the realism and utility of synthetic datasets for dynamic scenarios.

In addition, we plan to investigate the incorporation of state-of-the-art object detection architectures, such as YOLOv8, into the dual-task framework to assess their effectiveness and practicality in domain-specific applications. Furthermore, implementing the image generation pipeline in real time aboard ASVs has the potential to enable closed-loop learning systems that can continuously adapt to evolving environmental conditions and detection requirements.

CRediT authorship contribution statement

Fredy Barrientos-Espillco: Conceptualization, Investigation, Methodology, Formal analysis, Resources, Software, Visualization, Writing – original draft. **Gonzalo Pajares:** Formal analysis, Investigation, Validation, Supervision, Writing – review & editing. **José A. López-Orozco:** Data curation, Funding acquisition. **Eva Besada-Portas:** Writing – review & editing, Supervision, Funding acquisition, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by the Research Projects IA-GES-

BLOOM-CM (Y2020/TCS-6420) of the Synergic program of the Comunidad Autónoma de Madrid, SMART-BLOOMS (TED2021-130123B-I00) funded by the Spanish Ministry of Science and Innovation and the European Union NextGeneration, and INSERTION (PID2021-127648OB-C33) of the Knowledge Generation Programs of the Spanish Ministry of Science and Innovation. The first author, Fredy Barrientos-Espillco, is supported by a scholarship by PRONABEC, Ministry of Education of Peru.

Data availability

The authors are unable or have chosen not to specify which data has been used.

References

Ahn, Y.-H., Shanmugam, P., Ryu, J.-H., & Jeong, J.-C. (2006). Satellite detection of harmful algal bloom occurrences in Korean waters. *Harmful Algae, 5*(2), 213–231. https://doi.org/10.1016/j.hal.2005.07.007

Alhabeeb, S. K., & Al-Shargabi, A. A. (2024). Text-to-image synthesis with generative models: Methods, datasets, performance metrics, challenges, and future direction. *IEEE Access, 12*, 24412–24427. https://doi.org/10.1109/ACCESS.2024.3365043

An, R., Yang, S., Lu, M., Zhang, R., Zeng, K., Luo, Y., Cao, J., Liang, H., Chen, Y., She, Q., Zhang, S., & Zhang, W. (2024, November 18). MC-LLaVA: Multi-Concept Personalized Vision-Language Model. arXiv preprint https://arxiv.org/abs/2411.11706v3.

Barrientos-Espillco, F., Gascó, E., López-González, C. I., Gómez-Silva, M. J., & Pajares, G. (2023). Semantic segmentation based on Deep learning for the detection of Cyanobacterial Harmful Algal Blooms (CyanoHABs) using synthetic images. *Applied Soft Computing, 141*, Article 110315. https://doi.org/10.1016/j.asoc.2023.110315

Barrientos-Espillco, F., Gómez-Silva, M. J., Besada-Portas, E., & Pajares, G. (2024). Integration of object detection and semantic segmentation based on Convolutional Neural Networks for navigation and monitoring of cyanobacterial blooms in lentic water scenes. *Applied Soft Computing, 111849*. https://doi.org/10.1016/j.asoc.2024.111849

Besada-Portas, E., Risco-Martín, J. L., Esteban San Roman, S., Girón-Sierra, J. M., Pajares, G., & López-Orozco, J. A. (2023). Tecnologías habilitadoras para automatizar la monitorización de blooms de cianobacterias. 6–11. https://ruc.udc.es/dspace/handle/2183/33530.

Bie, F., Yang, Y., Zhou, Z., Ghanem, A., Zhang, M., Yao, Z., Wu, X., Holmes, C., Golnari, P., Clifton, D. A., He, Y., Tao, D., & Song, S. L. (2025). RenAIssance: A survey into AI text-to-image generation in the era of large model. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 47*(3), 2212–2231. https://doi.org/10.1109/TPAMI.2024.3522305

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint https://doi.org/10.48550/arXiv.2004.10934.

Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. In Y. Lechevallier & G. Saporta (Eds.), Proceedings of COMPSTAT'2010 (pp. 177–186). Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-2604-3_16.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Cannizzaro, J. P., Barnes, B. B., Hu, C., Corcoran, A. A., Hubbard, K. A., Muhlbach, E., Sharp, W. C., Brand, L. E., & Kelble, C. R. (2019). Remote detection of cyanobacteria blooms in an optically shallow subtropical lagoonal estuary using MODIS data. *Remote Sensing of Environment, 231*, Article 111227. https://doi.org/10.1016/j.rse.2019.111227

Chang, H., Zhang, H., Barber, J., Maschinot, A. J., Lezama, J., Jiang, L., Yang, M.-H., Murphy, K., Freeman, W. T., Rubinstein, M., Li, Y., & Krishnan, D. (2023). Muse: Text-To-Image Generation via Masked Generative Transformers. arXiv preprint https://doi.org/10.48550/arXiv.2301.00704.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), Computer Vision – ECCV 2018 (pp. 833–851). Springer International Publishing. https://doi.org/10.1007/978-3-030-01234-2_49.

Chen, N., Wang, S., Zhang, X., & Yang, S. (2020). A risk assessment method for remote sensing of cyanobacterial blooms in inland waters. *Science of The Total Environment, 740*, Article 140012. https://doi.org/10.1016/j.scitotenv.2020.140012

Dong, Z., Wei, P., & Lin, L. (2023). DreamArtist: Towards Controllable One-Shot Text-to-Image Generation via Positive-Negative Prompt-Tuning. arXiv preprint https://doi.org/10.48550/arXiv.2211.11337.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. arXiv preprint https://doi.org/10.48550/arXiv.2208.01618.

Graham, J. L., Dubrovsky, N. M., & Eberts, S. M. (2016). Cyanobacterial harmful algal blooms and U.S. Geological Survey science capabilities. In Cyanobacterial harmful algal blooms and U.S. Geological Survey science capabilities (USGS Numbered Series Nos. 2016–1174; Open-File Report, Vols. 2016–1174, p. 12). U.S. Geological Survey. https://doi.org/10.3133/ofr20161174.

Hamilton, D. P., Wood, S. A., Dietrich, D. R., & Puddick, J. (2014). Costs of harmful blooms of freshwater cyanobacteria. In Cyanobacteria (pp. 245–256). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118402238.ch15.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems, 33*, 6840–6851. https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Hu, C. (2009). A novel ocean color index to detect floating algae in the global oceans. *Remote Sensing of Environment, 113*(10), 2118–2129. https://doi.org/10.1016/j.rse.2009.05.012

Huisman, J., Codd, G. A., Paerl, H. W., Ibelings, B. W., Verspagen, J. M. H., & Visser, P. M. (2018). Cyanobacterial blooms. *Nature Reviews Microbiology, 16*(8), Article 8. https://doi.org/10.1038/s41579-018-0040-1

Jiang, J., Zhang, Y., Feng, K., Wu, X., Li, W., Pei, R., Li, F., & Zuo, W. (2024). MC^2: Multi-concept Guidance for Customized Multi-concept Generation. arXiv preprint https://doi.org/10.48550/arXiv.2404.05268.

Jones, M., Wang, S.-Y., Kumari, N., Bau, D., & Zhu, J.-Y. (2024). Customizing Text-to-Image Models with a Single Image Pair. In *SIGGRAPH Asia 2024 Conference Papers* (pp. 1–13). https://doi.org/10.1145/3680528.3687642

Kang, M., Zhu, J.-Y., Zhang, R., Park, J., Shechtman, E., Paris, S., & Park, T. (2023). Scaling Up GANs for Text-to-Image Synthesis. 10124–10134. https://openaccess.thecvf.com/content/CVPR2023/html/Kang_Scaling_Up_GANs_for_Text-to-Image_Synthesis_CVPR_2023_paper.html.

Kidder, B. L. (2024). Advanced image generation for cancer using diffusion models. *Biology Methods & Protocols, 9*(1), Article bpae062. https://doi.org/10.1093/biomethods/bpae062

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2304.02643

Kumari, N., Zhang, B., Zhang, R., Shechtman, E., & Zhu, J.-Y. (2023). Multi-concept customization of text-to-image diffusion. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023*, 1931–1941. https://doi.org/10.1109/CVPR52729.2023.00192

Kutser, T., Metsamaa, L., Strömbeck, N., & Vahtmäe, E. (2006). Monitoring cyanobacterial blooms by satellite remote sensing. *Estuarine, Coastal and Shelf Science, 67*(1), 303–312. https://doi.org/10.1016/j.ecss.2005.11.024

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., & Zhang, L. (2023). Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv preprint https://doi.org/10.48550/arXiv.2303.05499.

Liu, V., & Chilton, L. B. (2022). Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–23). https://doi.org/10.1145/3491102.3501825

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems, 36*, 46534–46594.

Mansimov, E., Parisotto, E., Ba, J. L., & Salakhutdinov, R. (2016). Generating images from captions with attention. arXiv preprint https://doi.org/10.48550/arXiv.1511.02793.

Nguyen, T. V., Glaser, A., & Biessmann, F. (2024). Generating synthetic satellite imagery with deep-learning text-to-image models—Technical challenges and implications for monitoring and verification. arXiv preprint https://doi.org/10.48550/arXiv.2404.07754.

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., & Chen, M. (2022). GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 16784–16804).

van den Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel Recurrent Neural Networks. In *Proceedings of The 33rd International Conference on Machine Learning* (pp. 1747–1756).

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., & Rombach, R. (2023). SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint https://doi.org/10.48550/arXiv.2307.01952.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022, April 13). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv.Org. https://arxiv.org/abs/2204.06125v1.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8821–8831).

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, 1060–1069.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022*, 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. 22500–22510. https://openaccess.thecvf.com/content/CVPR2023/

html/Ruiz_DreamBooth_Fine_Tuning_Text-to-Image_Diffusion_Models_for_Subject-Driven_Generation_CVPR_2023_paper.html.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems, 35*, 36479–36494.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2025). A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint https://doi.org/10.48550/arXiv.2402.07927.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., … Resnik, P. (2025). The prompt report: A systematic survey of prompt engineering techniques. arXiv preprint https://doi.org/10.48550/arXiv.2406.06608.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint https://doi.org/10.48550/arXiv.2307.09288.

Tran, G.-N., Che, Q.-H., Vu, T.-T. D., Pham, B.-N., Nguyen, V.-T., Le, T.-N., & Tran, M.-T. (2025). FaR: Enhancing multi-concept text-to-image diffusion via concept fusion and localized refinement. arXiv preprint https://doi.org/10.48550/arXiv.2504.03292.

Vatsal, S., & Dubey, H. (2024). A survey of prompt engineering methods in large language models for different NLP tasks. arXiv preprint https://doi.org/10.48550/arXiv.2407.12994.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Liu, S., Berman, W., Xu, Y., & Wolf, T. (2024). Diffusers: State-of-the-art diffusion models (Version 0.12.1) [Python]. https://github.com/huggingface/diffusers (Original work published 2022).

Wang, Z. M., Peng, Z., Que, H., Liu, J., Zhou, W., Wu, Y., Guo, H., Gan, R., Ni, Z., Yang, J., Zhang, M., Zhang, Z., Ouyang, W., Xu, K., Huang, S. W., Fu, J., & Peng, J. (2024). RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. arXiv preprint https://doi.org/10.48550/arXiv.2310.00746.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), Computer Vision – ECCV 2018 (pp. 3–19). Springer International Publishing. https://doi.org/10.1007/978-3-030-01234-2_1.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. 1316–1324. https://openaccess.thecvf.com/content_cvpr_2018/html/Xu_AttnGAN_Fine-Grained_Text_CVPR_2018_paper.html.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems, 36*, 11809–11822.

Yao, Z., Feng, F., Li, R., & Wang, X. (2024). Concept conductor: Orchestrating multiple personalized concepts in text-to-image synthesis. arXiv preprint https://doi.org/10.48550/arXiv.2408.03632.

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., & Sang, N. (2018). BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), Computer Vision – ECCV 2018 (pp. 334–349). Springer International Publishing. https://doi.org/10.1007/978-3-030-01261-8_20.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2017). StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. 5907–5915. https://openaccess.thecvf.com/content_iccv_2017/html/Zhang_StackGAN_Text_to_ICCV_2017_paper.html.

Zhang, N., & Tang, H. (2024). Text-to-image synthesis: A decade survey. arXiv preprint https://doi.org/10.48550/arXiv.2411.16164.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017*, 6230–6239. https://doi.org/10.1109/CVPR.2017.660