Generative AI methods for synthesis of image data to train AI for automated scene understanding in a military context: a review of opportunities

Ella P. Fokkinga[a], Thijs A. Eker[a], Jan Erik van Woerden[a], Jean-Michel Witon[b], Simon O.B. Stallinga[a,c], Arnoud Visser[c], Klamer Schutte[a], and Friso G. Heslinga[a]

[a]TNO - Intelligent Imaging, Oude Waalsdorperweg 63, the Hague, the Netherlands
[b]John Cockerill Defense, 57925 Distroff, France
[c]University of Amsterdam, the Netherlands

ABSTRACT

The rapid increase in sensors on manned and unmanned military platforms has led to a significant rise in image data (e.g., visible, infrared, sonar, radar), enabling extensive scene analysis. Thorough and real-time understanding of these scenes requires automated image analysis tools, for e.g. object detection, traversability analysis, and threat classification. However, the development of artificial intelligence (AI) models for automated scene understanding is constrained by limited access to relevant military training data due to its restricted nature, high acquisition costs, and evolving threat signatures. Several studies highlight the potential of synthetic data as an alternative to measured training data, for example by utilizing physics-based modeling of scenes and objects of interest.

Recent advances in generative AI (GenAI), particularly in diffusion-based models, offer opportunities to synthesize data with variations beyond what was previously possible, improving performance in various non-military image analysis tasks. Despite this, the lack of military-relevant data used for GenAI model development suggests that non-specialized models may produce military scenes with limited quality and variation. In this review, we explore the opportunities of state-of-the-art GenAI methods for creating high-quality training data for military AI systems. We identify three key strategies: (1) full-image generation by fine-tuning with application-specific data; (2) inpainting, where objects of interest can be placed in existing image data; and (3) image-to-image translation which is used to augment image conditions or translate between image modalities. Visual results of each of these methods are promising. Some studies have already shown benefits of these data synthesis methods as data augmentation to improve downstream AI models. Further research shall determine the value for operationalization in a wide set of use-cases.

Keywords: Generative AI; Data synthesis; Deep learning; Situational awareness; Scene understanding; Object detection

1. INTRODUCTION

The modern battlefield is increasingly characterized by its transparency, a phenomenon driven by the rapid proliferation of sensors. These sensors, which encompass surveillance, intelligence, and reconnaissance (ISR) systems, as well as military platforms, both manned and unmanned, have led to a substantial increase in the volume of image data. This surge in data requires automated processing and analysis techniques to effectively utilize the information gathered. Typical image data encompasses visible-light (RGB) and infrared (IR) imagery. However, one can also consider other image data modalities, such as Synthetic Aperture Radar (SAR) and Synthetic Aperture Sonar (SAS). All of these types of data must be processed into relevant information to achieve a thorough and real-time understanding of the scenes, necessitating the use of automated image analysis tools. These tools include object detection,[1,2] traversability analysis,[3,4] and threat classification,[5,6] and are referred to as *downstream tasks*.

Corresponding author: Friso G. Heslinga. E-mail: fgheslinga@gmail.com

The development of artificial intelligence (AI) models for automated scene understanding is constrained by limited access to relevant military training data due to its restricted nature, high acquisition costs, and evolving threat signatures. Several studies highlight the potential of synthetic data as an alternative to measured training data. A typical strategy involves utilizing 3D-based simulations of scenes and objects of interest, followed by the acquisition of various viewpoints by modeling the relevant sensor. This can be achieved using publicly available simulation software for the generation of RGB data.[7–10] Similar work is also done to simulate IR-imagery,[11] where thermal properties can be modeled explicitly.[12] The studies demonstrate that adding synthetic data as training data improves performance on downstream tasks, particularly when the number of real samples is low. For other applications, especially non-RGB, custom-designed simulation software is required.[13,14] Although the results of AI models developed with simulated data are promising, they come at the cost of long simulator development times, depend on the quality and detail of the 3D models used,[15] and the amount of variability available in the simulations.[7] Adaptability to new settings is possible, but requires extensions to the simulator, which can be costly and time-consuming.

Recent advances in generative AI (GenAI) offer opportunities to synthesize image data with variations beyond what was previously possible. Some of the published results look realistic and include a wide range of variation in composition and style.[16,17] Not only do the images look visually impressive, they have been used to extend training data for improving performance in various non-military image analysis tasks. For example, Azizi et al. (2023)[18] achieved state-of-the-art image classification performance for the ImageNet benchmark dataset by adding a large amount of images synthesized with GenAI. In a similar fashion, Du et al. (2023)[19] generated photorealistic outliers to supplement ImageNet training data to successfully adapt classification to out-of-domain datasets.

Similarly, GenAI is likely to enable the synthesis of training data for AI applications in the military domain. For example, by creating a large set of images with military vehicles. However, off-the-shelf GenAI models lack knowledge of the military domain, since they are not trained with many relevant training examples. To create image data relevant for military applications, extra steps are often needed. These may include fine-tuning models on domain-specific examples or synthesizing only parts of the image, such as the background or specific objects, thereby reducing the need for extensive military knowledge. In this review, we explore the opportunities of state-of-the-art GenAI methods to create high-quality training data for military AI systems, mainly focusing on RGB and IR image data.

1.1 Related surveys and study contribution

GenAI methods for image synthesis have been widely studied in recent years, with extensive surveys thoroughly analyzing their development and applications across various domains, including medical imaging, entertainment or art.[20] However, studies that consider their potential for military AI are lacking.

Several studies have provided comprehensive overviews of generative models for image synthesis, but typically focus on only one class of generative models, such as Generative Adversarial Network (GAN)-based approaches. For example, Wang et al. (2020)[21] presented a state-of-the-art review on various GAN-based image synthesis tasks, while Alotaibi (2020)[22] focused specifically on GANs for image-to-image translation, covering both supervised and unsupervised approaches. Expanding on these, Shamsolmoali et al. (2021)[23] reviewed GANs across multiple synthesis tasks, including image-to-image translation and text-to-image generation. Alhabeeb et al. (2024)[24] provide an overview of text-to-image generation using GANs.

More recent surveys focus on diffusion-based methods, such as Zhan et al. (2024),[25] who examine the role of diffusion models for seven conditional image synthesis tasks, including text-to-image, image editing, and visual signals (such as a sketch) to image translation. Zhang et al. (2024)[26] provide a comprehensive review on diffusion-based text-to-image models.

Baraheem et al. (2023)[27] provide a broad and structured review on image synthesis techniques, considering several model types (GANs, Variational Autoencoders (VAE), diffusion models) and input modalities. However, this survey, like the others, focuses on image synthesis as a standalone task and does not consider the use of synthetic images to train AI models. Other surveys, such as by Alzubaidi et al. (2023),[28] do consider generative AI as a tool for data augmentation. This survey is however a broader review of deep learning strategies to address data scarcity, not specifically focused on image synthesis.

In this review, we aim to bridge this gap by:

1. Providing a broad perspective that covers multiple image synthesis models: GANs, VAEs, diffusion, and hybrid models;
2. Examining not only the synthesis techniques but also how they could be used to support training of AI-models;
3. Analyzing the potential of the methods for military applications, where the available real-world data is limited.

This review provides a comprehensive overview of GenAI methods for image synthesis, with a particular focus on their potential for AI model training in military applications. The paper is organized as follows: in Section 2, we start with an introduction to the three most widely used generative models, GANs, VAEs, and diffusion models, establishing a technical foundation for understanding their role in image synthesis. Next, in Section 3, the review categorizes image synthesis tasks into three main groups: 1) full image generation, 2) in-painting, 3) image-to-image translation. For each category, we examine relevant research papers, including studies that explicitly evaluate their models for AI training (data augmentation) but also those that treat image synthesis as a standalone task. To illustrate the potential of these methods, from each category, we include a practical example where we generate synthetic images for military-related use cases. We conclude with an overall discussion in Section 4, where we describe where the greatest potential lies for military AI applications and the key challenges that must be addressed.

2. MODELS

2.1 VAE

Kingma and Welling introduced the VAE in 2013.[29] A VAE consists of an encoder-decoder architecture: the encoder maps input images x to a latent distribution parametrized by mean μ and variance σ^2, improving upon regular Autoencoders, which directly map to a single point in the latent space. Consecutively, a latent variable z is sampled from the learned distribution and is mapped back to the original data space, producing an image $\tilde{x}$. During training, the loss consists of two components: the reconstruction loss, which ensures that the image $\tilde{x}$ generated by the decoder closely resembles the input image x, and the Kullback-Leibler (KL) divergence loss, which regularizes the latent space so that it allows for meaningful samples. This ensures that the learned latent space is continuous, allowing smooth interpolation between data points.[20] After training the VAE, instead of first encoding an image and then decoding it, we can directly sample latent variables from the prior distribution, i.e. the sampled z is passed through the decoder to generate synthetic images. VAEs have been applied for various image synthesis tasks, for instance to generate faces, medical images, or satellite imagery. Baraheem et al. (2023)[27] published an overview of VAE-based methods for text-to-image generation in 2023.

2.2 GAN

The GAN was initially proposed by Goodfellow et al. in 2014.[30] Since then, GANs have been widely used for numerous image processing tasks.[20–22,24,28,31] The model consists of a generator G and a discriminator D. The generator attempts to create realistic images $\tilde{x}$ from a random noise distribution p_z from which we can sample a latent variable z. The discriminator acts as a binary classifier, trying to distinguish generated images $\tilde{x}$ from real images x. D and G are alternately updated, where D is being trained to maximize the probability of correctly classifying real and generated images, while G is trained to "fool" the discriminator D so that images are incorrectly classified. Both the generator and the discriminator can be any type of network.[22] After training, to generate images, a random noise vector z is sampled from the noise distribution p_z and passed by G to generate a realistic image $\tilde{x}$, i.e. the synthetic image.

Examples of applying GANs to improve military datasets are scarce, but do exist. McCloskey et al. (2023)[32] use different GAN variants to augment training data for an ISR-object detector. Wang et al. (2024)[33] propose a data augmentation method for conditional inpainting of multiple object classes in IR images of a naval setting. Koo et al. (2024)[34] developed a generator that produces synthetic sonar images, used to augment training data for an underwater classification model. In the review by Zhang et al. (2023),[35] GAN-based methods to generate Synthetic Aperture Radar (SAR) and/or inverse-SAR imagery are described.

2.3 Diffusion

The most recent developments in the field of image synthesis have taken place using Diffusion models, obtaining an improved image realism in comparison with GANs.[24,36] Diffusion models consist of two phases: forward and backward diffusion. In the forward phase, Gaussian noise is added to the input image x step by step. After T steps, the image is nearly indistinguishable from pure Gaussian noise. In the backward phase, the diffusion process is reversed, where the noise is iteratively removed so that the original input image is recovered. During training, the model has to predict the noise η at time step t. As loss, the predicted noise is compared to the true noise. After training, a random noise vector can be sampled and by applying the learned reverse process from T to 0 to this noise vector, finally a synthetic image x_0 is generated. Diffusion models can be applied for a variety of image synthesis tasks, including full image generation, inpainting and image-to-image translation.[24,36,37]

3. METHODS FOR IMAGE SYNTHESIS WITH GENAI

GenAI can be used in various ways for creating synthetic image data. In this chapter, we review specific methods, providing an overview of the current playing field, and focus on methods that we consider to have most potential at this point in time. Methods are grouped into three main strategies. The first one is full-image generation, where the goal is to synthesize a full image, e.g. by finetuning a diffusion model with some domain-specific examples. The next strategy is inpainting, where parts of an image are filled in, for example to place objects in existing images that represent the environment of interest. The final set of methods is about image-to-image translation. Here, the main goals are to translate either the appearance of the image (e.g. change weather conditions), or the modality (e.g. translate RGB tot IR data).

An overview of our mapping of methods is provided in Fig. 1. Each GenAI strategy is subdivided into some subcategories with some examples of key papers on this topic. In the next sections, we dive into the different strategies. For most papers, we only explore the opportunities of synthesizing image data. Whenever available however, we add some context on how the results can be used to augment training data and improve downstream AI model development.

3.1 Full-image generation

Recent advances in diffusion models have established them as the state-of-the-art for high-quality image generation. Well-known diffusion models, such as Stable Diffusion[38] and Imagen,[39] are trained on large-scale general-purpose data sets such as LAION-2B and LAION-5B,[40,41] which contain approximately 2 to 5 billion image-text pairs collected from the web. Although this allows them to generate photorealistic images across a wide range of domains, they lack specific knowledge in the military domain. This knowledge gap means that out-of-the-box models are likely to be insufficient for military applications. However, fine-tuning techniques can be used to adapt the models. Furthermore, the image synthesis can be controlled by specifying the prompts and incorporating guidance methods.

The rest of this section is thus structured as follows: first, we introduce the major base diffusion models, followed by an overview of fine-tuning methods. Next, prompting strategies to guide the image generation are described. Finally, we discuss additional guidance techniques, such as ControlNet[42] and other conditioning mechanisms, to further enhance control over the generated images.

3.1.1 Base models

To reduce computational costs, modern diffusion models rely on compressed latent spaces rather than performing the diffusion process directly in pixel space. Instead of generating images at full resolution, diffusion models first encode images into a lower-dimensional latent space, where the generative process is performed before decoding the results back into high-resolution images. This is typically achieved using VAEs (Section 2.1), or Vector Quantized GANs (VQGAN).[43]

VAEs encode images into a continuous latent space, enabling smooth transformations but sometimes introducing blurriness, while VQ-GANs quantize images into discrete tokens, preserving sharpness and structure but occasionally causing visual artifacts due to reconstruction limitations. Both of these approaches significantly reduce computational requirements, making training and inference more efficient. A downside, however,

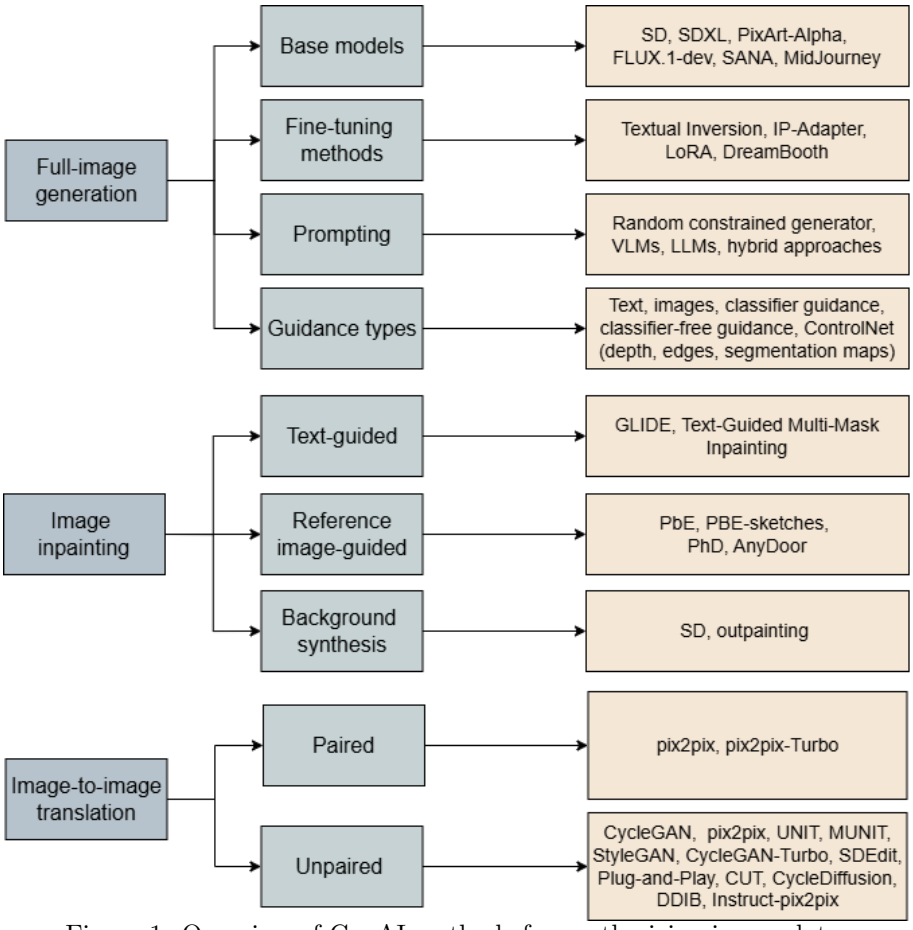

Figure 1: Overview of GenAI methods for synthesizing image data.

of incorporating these sub-models, is that the final image quality depends on the effectiveness of the VAE or VQ-GAN for encoding and decoding. Most modern latent diffusion models, including Stable Diffusion (SD), use VAEs rather than VQ-GANs due to their better adaptability for text-to-image generation and fine-tuning flexibility.[44] In the next paragraphs, we describe the key base diffusion models, their architectures, and their specific advantages for image synthesis.

Stable Diffusion 1.5 Following OpenAI's success with DALL-E, which was based on a discrete VAE,[17] SD was introduced. SD emerged from the Latent Diffusion Model (LDM) framework,[38] combining a U-Net-based CNN architecture with transformer cross-attention blocks at each layer. One of the major impacts of SD1.5 was due to its open-source nature. The open release of the model's weights led to widespread adoption in artistic communities, and contributed to the development of finetuning tools for the base model. This adaptability makes it a good candidate for customization, including for military applications. However, one limitation of the model is its training and output resolution of 512×512 pixels, which does not preserve finer details as well as more modern models with a larger output size. Additionally, a Contrastive Language-Image Pretraining (CLIP)-based encoder[45] is used, which is restricted to a using relatively short prompt length. For more details on the CLIP encoder used in SD, refer to Section 3.1.3.

SDXL Released in 2023, Stable Diffusion XL (SDXL)[44] significantly improved upon SD1.5, enhancing image quality, resolution, and detail preservation. SDXL leverages a three times larger U-Net, a two-stage synthesis pipeline in latent space, and an improved VAE for higher fidelity decoding. In addition, a larger CLIP model

is used for text encoding, enabling better text-image alignment. The training procedure was improved with multi-aspect ratio learning and a cropping mechanism, allowing the model to generalize across diverse image dimensions and to retain finer details. Furthermore, the model is able to synthesize images up to 1024×1024 pixels, making it more suitable for detailed scene synthesis.

PixArt-Alpha PixArt-Alpha introduced Diffusion Transformers (DiTs)[46] as a replacement for the U-Net architecture found in SD1.5 and SDXL. The authors demonstrated that Vision Transformers (ViTs) can be effectively adapted for image generation when scaling both model parameters and training dataset size is feasible. Despite this shift in network architecture, PixArt-Alpha retains SD's VAE to decode from the latent space back into pixel space. In this way, compatibility is ensured with existing fine-tuning methods like LoRA and DreamBooth (Section 3.1.2). This allows for experimentation with the DiT architecture while leveraging the existing SD ecosystem. Additionally, PixArt-Alpha integrates the Vision-Language Model (VLM) LLaVA[47] to generate synthetic captions for training images, enhancing text-to-image alignment during training.

FLUX.1-dev FLUX.1-dev[16] is a new diffusion model developed by the original authors of VQ-GAN,[43] LDM, and Stable Diffusion.[38] Black Forest Lab, the company behind its development, has positioned it as a commercial alternative to SD. While no formal research paper accompanied its release, a blog post hints at its architecture, notably the use of Rectified Flow Transformers (RFTs).[48] While DiTs, used in PixArt-Alpha, rely on complex, multi-step diffusion processes, RFTs connect data and noise in a more direct trajectory, simplifying the generation process. This approach reduces sampling steps, leading to faster generation times and improved image quality. Additionally, FLUX.1-dev includes native support for LoRA fine-tuning (Section 3.1.2, making it highly adaptable to a new subject or style in the military domain.

SANA SANA is a highly efficient diffusion model designed to generate high-resolution images up to 4096×4096, developed by NVIDIA and MIT researchers.[49] It integrates linear attention instead of self-attention for faster inference, and the small decoder-only Large Language Model (LLM) Gemma-2 as a text encoder, to enhance text-to-image alignment. In addition, an improved solver (Flow-DPM-Solver) reduces the number of sampling steps, making inference up to $100\times$ faster than FLUX and $40\times$ for SDXL.

3.1.2 Fine-tuning methods

In less than six months after the release of SD,[38] several methodologies were introduced to adapt text-to-image models to new subjects, styles, or datasets. Fine-tuning techniques allow models to incorporate domain-specific knowledge without requiring full retraining. In this section, we explore four major fine-tuning methods: Textual Inversion,[50] IP-Adapter,[51] DreamBooth,[52] and LoRA.[53]

Textual Inversion[50] learns a compact textual embedding from only a few reference images, allowing the model to reproduce the learned concept via prompting without modifying its weights. While computationally efficient, this method lacks structural control, making it less effective for precise scene modifications needed in military applications.

IP-Adapter[51] introduces a lightweight image-conditioned adapter module that allows a diffusion model to incorporate reference images into its generation process. Unlike Textual Inversion, which operates through text embeddings, IP-Adapter injects reference image features into the model's conditioning process. While this enables style and structural adaptation, it does not alter the model's weights,

DreamBooth[52] fine-tunes a model using a small dataset and a unique abstract identified (e.g. "$[V]$") representing the object or style. This method maintains the object's key visual characteristics and allows for synthesis of the object on new backgrounds. For instance, if fine-tuned on images of a CV90 armored vehicle, the model could generate realistic CV90 images in different landscapes and conditions. However, DreamBooth fine-tunes all model weights, producing a new model that matches the original in storage size, which makes it impractical for fine-tuning on a large variety of military vehicles or environmental conditions.

LoRA (Low-Rank Adaptation)[53] was initially developed for LLM fine-tuning, but has since then been adapted for diffusion models. Instead of modifying the entire model, LoRA injects small, trainable low-rank matrices

into specific layers (typically the attention layers), while keeping the original weights frozen. This drastically reduces the computational cost to adapt to a new object as well as the storage requirements, making LoRA ideal for military applications that require rapid, efficient adaptation. Another major advantage of LoRA is its ability to combine multiple fine-tuned adaptations. A single model can be merged with several LoRAs, allowing for separate tuning of content (e.g. specific military vehicle types) and style (e.g. weather conditions, pseudo-thermal imagery, sensor noise). However, in prelimimary work, we found that performance can degrade when too many LoRAs are merged. In Fig. 2, we show some visual results of FLUX models finetuned with LoRA.

Figure 2: Visual examples of images generated with FLUX.1-dev finetuned with LoRA on an in-house dataset. Left and center images are created with a model fine-tuned on a set of RGB photos of military vehicles. Right example is a result from a model fine-tuned on long wave infrared images.

Among these methods, LoRA stands out as the most practical for military applications, particularly for modern DiT-based models, due to its efficiency, flexibility, and ability to combine multiple fine-tuned adaptations. While DreamBooth provides high-fidelity subject learning, the large storage size required makes it less scalable for diverse military datasets. IP-Adapter and Textual Inversion offer lighter-weight alternatives, but their limited structural control reduces their suitability for complex scene adaptation.

3.1.3 Prompting

A key component of diffusion models is their text encoder, which maps prompts into a shared latent space to align text and images. The prompt is the main input for the different families of text-to-image diffusion models and thus directly influences the generated output. Two common approaches to synthetically generate prompts are via a constrained random generator, and via the use of VLMs and LLMs.

A constrained random generator constructs prompts using a predefined set of attributes or keywords, that are structured to form a coherent sentence. Attributes may include the points of view, objects (or identifiers for fine-tuned models, see Section 3.1.2, DreamBooth), backgrounds, times of day, and weather conditions. Each attribute is assigned a task-specific vocabulary. The prompt generator randomly selects a value from this vocabulary for each attribute, which can be used directly at runtime, or pre-generated as part of a "prompts bank".

An alternative approach to prompt generation is the use of VLMs and LLMs, which can automatically generate or refine prompts based on textual or visual inputs. VLMs can analyze existing images and generate captions or descriptions of details.[47] Unlike constrained random generators, VLMs are steerable via system prompts, allowing users to define how images should be described. In the military domain, this ability could be particularly useful for specialized imagery, such as IR-images. By instructing a VLM to reason about how different materials radiate or reflect heat, it can generate structured descriptions of object appearances in shades of gray.

Finally, random prompt generators can be combined with VLM-generated captions or LLM refinements to enhance diversity. For example, a random generator may produce a generic prompt (e.g., "A military vehicle

in a desert environment"), which a VLM then enriches with contextual details (e.g., "The armored vehicle is kicking up sand, with heat distortion visible in the background"). An LLM can further refine or merge prompts from multiple images, creating more varied and realistic synthetic training data.

As mentioned in Section 3.1.1, models like SD1.5 rely on CLIP-based encoders,[45] which are restricted to 77 tokens (or $\sim$ 50-60 words), with an effective length of only $\sim$ 20 tokens - likely constrained by the alt-text dataset used for its training.[54] This restriction results in CLIP-based models performing best with short, keyword-driven prompts over detailed natural language description. To address this, newer architectures replace or extend CLIP with LLMs to allow longer, richer, and more nuanced prompts. For example, FLUX-1.dev[16] incorporates a T5-XXL encoder-decoder, while SANA[49] utilizes a Gemma-2-2B decoder-only model for improved text encoding, increasing the supported prompt length to several hundred or even over 1,000 tokens, depending on the model configuration.

3.1.4 Guidance beyond text prompts

Diffusion models can be guided beyond simple text prompts through various techniques that influence the generative process. The most straightforward is the previously described text-to-image approach, where the model generates an image purely from a textual input prompt. The output quality is highly dependent on the quality of the prompt, requiring prompts that align closely with the model's pretraining or fine-tuning data.

A second option is using image-to-image, where an input image along with a prompt is used to generate a new image. Instead of starting from pure random noise, the model adds noise to the original image and gradually refines it, allowing to steer the output of the model to a different region of the latent space.

More structured forms of guidance have also been proposed. One such method is classifier guidance,[55] where an external classifier is trained to predict class labels on noisy images. During sampling, the gradient of this classifier's output is used to steer the diffusion model's denoising trajectory toward a desired class. While effective, this method requires a separate classifier trained on noisy data and can sometimes lead to overconfident or less diverse outputs. To address this, classifier-free guidance[56] removes the need for an external model by training the diffusion model to handle both conditional (with prompts) and unconditional (with empty prompts) cases. At inference, the model's own conditional and unconditional outputs are combined to guide generation, striking a balance between sample quality and diversity. This technique has become the standard for modern text-to-image models, including SD (Section 3.1.1), due to its simplicity and strong performance.

Finally, ControlNet enables additional conditioning[42] on structured signals beyond text or class labels. While alternative approaches exist, such as the semantic layout-guided method described by Wang et al. (2022),[57] ControlNet is particularly notable for its general-purpose design, modularity, and flexibility across diverse control types. It supports various structural priors such as depth maps, Canny edge maps, and segmentation masks, which can be extracted from arbitrary images. These signals act as structural constraints, helping to preserve object shape or enforce specific scene layouts. For instance, depth and edge maps help maintain object structures, while segmentation masks allow explicit control over object placement and composition in the generated output. A visual example of this capability is provided in Fig. 3, where a depth map is employed to guide an SDXL model (finetuned with LoRA) via ControlNet.

While structured guidance methods such as ControlNet and image-to-image condition allow for controlled scene adaptation, another approach to generate various environmental conditions is through prompt engineering in closed-source models like MidJourney. Rothmeier et al. (2024)[58] use MidJourney to generate adverse weather images and show that fine-tuning their automotive object detector on this generated data improves detection performance. However, unlike SD, FLUX-1.dev, and SANA, MidJourney does not support direct fine-tuning, making it less suitable for generation of military-specific datasets.

3.1.5 Non-EO image synthesis

Given that the majority of base models for full-image generation are foundation models trained on extensive RGB datasets, it is anticipated that the predominant advancements will occur within this modality. Nevertheless, as can be seen in Fig. 2, finetuning a base model with non-RGB data can be used to synthesize pseudo-thermal infrared images. Synthesizing non-EO images (e.g. SAR/SAS) with this method might be more challenging since the appearance is very different. A recent survey by Huang et al. (2025)[59] shows the poor performance

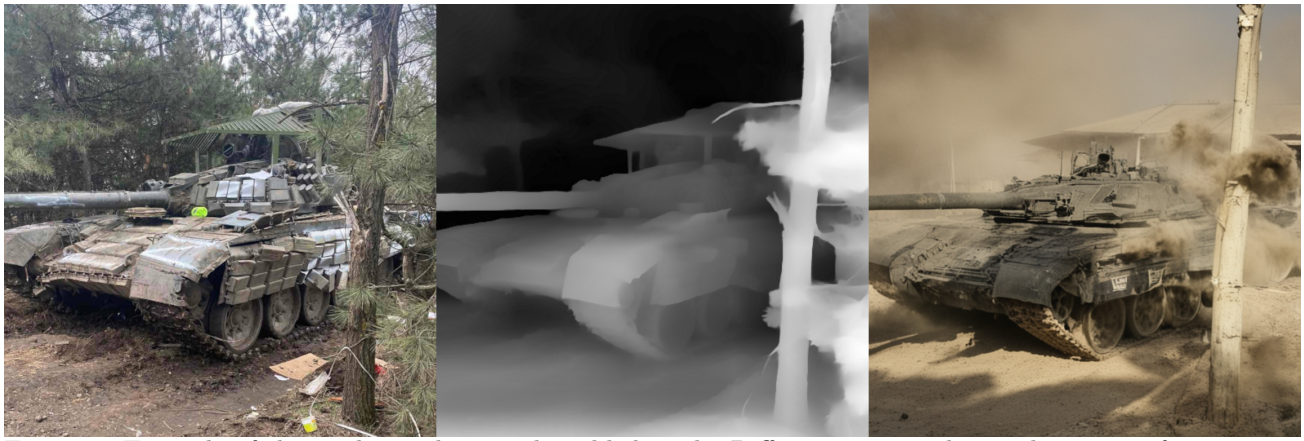

Figure 3: Example of the guidance that can be added to the Diffusion process, by combining a reference map with ControlNet. The center image represents a depth map of the photo on the left. The depth map is used to guide a SDXL model (finetuned with LoRA on RGB photos of military vehicles) to generate an image with a pose that is very similar to the photo on the left.

of out-of-the-box diffusion models for synthesizing SAR data, but also describes alternatives. For example, Song et al. (2022)[60] proposed an adversarial autoencoder to generate SAR images with aspect angular diversity to improve performance on a moving and stationary target recognition dataset. Similarly, Ju et al. (2024)[61] used a cascaded GAN (CGAN) to gradually improve the quality of the generated SAR ship images, increasing detection performance of SAR ship targets. We also identified advances of synthesizing data with GenAI for sonar imagery. For example, Koo et al. (2024)[34] leverage a CycleGAN to synthesize underwater sonar images to improve training of a target classifier, while Zhang et al. (2024)[62] managed to train a diffusion-based model to generate side-scan sonar images.

3.2 Inpainting

One disadvantage of full-image generation is the need for domain knowledge on all aspects of the image, including the foreground, background, and style. An alternative is to use inpainting, where only parts of the image are filled in. For example, by including an extra object of interest at a masked location, or by including some mission-specific clutter. Since less context is needed for inpainting, out-of-the-box models might require no finetuning, or less finetuning than full-image generation models.

This section covers two inpainting methods: text-based inpainting, which uses textual descriptions to guide the restoration process, and reference image-based inpainting, which leverages additional images to fill in gaps. We conclude this section with a brief discussion of outpainting, a technique for generating backgrounds.

3.2.1 Text-based inpainting

As described in Section 3.1, text-to-image models enable generation of multiple image samples from text prompts. Text-to-image inpainting models offer precise control over pose and location using masks, which is important for augmenting object detection datasets. By creatively optimizing prompts, engineers can produce diverse and practical data without the need for additional real images.

While initial text-based inpainting relied heavily on GANs, such approaches remain relevant, particularly in structured or constrained augmentation tasks. For instance, Wang et al. (2024)[33] introduce DOCI-GAN, a detection-oriented conditional inpainting GAN that generates synthetic IR object images from bounding box annotations, treating inpainting as a mask-to-image translation task. Modern inpainting techniques are largely driven by diffusion models, especially those built on the SD framework[38] (Section 3.1.1), which iteratively refine images to produce high-quality, realistic outputs. These methods generally start with a pretrained foundation model and fine-tune it to be conditioned by both text prompts and image masks.

In 2021, Nichol et al.[63] showed that text-to-image models are capable of using inpainting to realistically create diverse images, focusing on the difference between guided diffusion and classifier-free guidance, showing

the superiority of the latter. More recently, Fanelli et al. (2025)[64] leveraged advancements in multi-modal LLMs to develop a dataset featuring multi-mask and paired prompts. This dataset serves as input for fine-tuning SD models specifically for multi-mask inpainting tasks, further improving image synthesis by integrating detailed text-driven context into the inpainting process.

Although the advances in the field are promising, data augmentation for military AI using these models remains challenging. Often, the models have been trained on datasets with limited military vehicles and realistic data such as low-resolution targets and adverse weather conditions (rain, fog). Fine-tuning models like these requires generating relevant prompts, which may also lack precision within the military domain, as LLMs struggle with details of lesser-known military vehicle types.

3.2.2 Reference image-based inpainting

(a) The dotted red line outlines the area for inpainting, while the continuous red rectangle highlights the reference image used as input.

(b) The AnyDoor model fills in the masked area with a military vehicle, however, it does not retain the vehicle features perfectly.

Figure 4: A qualitative example of zero-shot performance by AnyDoor on out-of-domain image data.

Reference image-based inpainting presents a promising approach, particularly suited to the military domain, where the complexity and specificity of objects such as military vehicles demand precise representation. Unlike text-based inpainting, which relies on descriptive prompts that often fall short of capturing the nuanced details of such vehicles, reference image-based inpainting uses an image as the foreground object. This could enable several critical tasks: accurately capturing the shape and texture of the object to maintain its identity, generating a transformed view of the object (e.g. changes in pose, size, and illumination), and altering the surrounding area to ensure a realistic integration of the object into the scene, such as adding appropriate shadows or track marks. The integrity of the inpainted images is crucial, as inaccuracies such as hallucinated features or incorrect elements could significantly impair the performance of downstream object detection models trained on these augmented images. Comparable needs for realism and control exist in other sensitive domains, such as medicine, where both image-based and text-based inpainting have already been successfully applied to enhance dataset diversity. Liu et al. (2025)[65] demonstrate the effectiveness of reference image-based inpainting by using real tumor images to guide a diffusion model in generating synthetic image-mask pairs, improving training data for tumor segmentation models. Pérez-García et al. (2024)[66] apply text-conditioned inpainting to chest X-rays to add or remove pathological features. This approach enables the creation of controlled dataset shifts to stress-test biomedical vision models, revealing biases and failure modes before deployment.

One of the first works using diffusion models for inpainting, Paint by Example (PbE) by Yang et at. (2022),[67] adapts a pre-trained SD text-to-image model by replacing the CLIP text encoder[45] (Section 3.1.3) with a CLIP image encoder. Key enhancements include a content bottleneck, which limits the amount of detail extracted from the reference image, preventing simple copying and ensuring essential content guides the inpainting. Augmentation techniques diversify features and mask shapes, while classifier-free sampling refines the output to align with the reference image's style and class. The pre-trained SD model is fine-tuned on the OpenImages dataset. While PbE successfully blends the reference image into its background, it struggles to maintain the precise features of the reference object, often producing an object of the same class and color but not the ex-

act features. This limitation is significant for applications requiring precise object recognition, such as military vehicle identification.

Building upon the PbE framework, Kim et al. (2023)[68] introduce the use of sketches in addition to reference images for image inpainting, which enhances user control over the editing process. This approach incorporates a sketch of the masked region into the diffusion process. The model is fine-tuned on a cartoon dataset, demonstrating that the addition of sketch lines can significantly help in maintaining the distinct characteristics of the reference image, suggesting a potentially valuable modification for tasks requiring precise identity preservation in data augmentation.

In concurrent work titled Paste, Inpaint and Harmonize via Denoising (PhD)[69] (2023), the authors introduce a novel image editing framework that significantly modifies the process used in previous methods like PbE. PhD employs a two-step process: initially, it uses an off-the-shelf segmentation model to extract and paste the subject from a reference image directly into a background scene. Subsequently, it utilizes a diffusion model, conditioned using a ControlNet-style[42] encoder that encodes the pasted image and a text prompt, to regenerate the original image. This approach maintains the pretrained diffusion model in a frozen state, allowing for faster training times and leveraging the model's robust synthesis capabilities without retraining. Notably, PhD appears to maintain object identity more effectively than PbE, which could be especially beneficial for applications requiring high fidelity in object representation. However, the use of a frozen core model might limit adaptability to highly specialized domains such as military imagery.

The AnyDoor framework,[70] also proposed in 2023, introduces multiple advancements in reference image-based inpainting, significantly improving identity preservation and customization. Unlike prior methods that rely on augmenting masks and reference images during training, AnyDoor leverages video data to extract realistic variations in pose, lighting, and perspective by sampling and transferring objects between video frames. This approach results in more diverse and realistic object transformations. Key innovations include using a segmentation model to remove background from reference images, enhancing object focus, and replacing the CLIP image encoder with the more robust DINOv2[71] encoder to generate object identity vectors. To compensate for the low-resolution output (16×16) of DINOv2, AnyDoor integrates high-frequency information (e.g., edges and patterns) through a ControlNet-style module,[42] ensuring the model captures fine-grained details essential for identity preservation. This makes it a promising tool for applications that require accurate and consistent object representation, especially in scenarios that require detailed fidelity. An example of the out-of-the-box performance of AnyDoor for inpainting a military relevant object is shown Fig. 4. Although AnyDoor excels in identity preservation, this example also demonstrates that handling out-of-domain data still presents challenges.

3.2.3 Background synthesis

A topic related to inpainting is background synthesis, where the goal is generate relevant background images. Prior to using GenAI, one strategy for using realistic backgrounds in simulation was the use of High Dynamic Range Imaging (HDRI) backgrounds, as demonstrated by Eker et al. (2023).[7] However, this method is limited in background variation and does not account for potential inconsistencies in appearance between foreground and background. Recent work has explored diffusion models for generating more diverse and adaptable backgrounds.

Pichler and Hueber (2024)[72] use a diffusion model (SD) to create full images as backgrounds, to combine with drone-captured vehicle footage, creating a large dataset for vehicle detection. However, their approach directly overlays objects on the background without any harmonization techniques, potentially leading to visual inconsistencies. Li et al. (2024)[73] consider background synthesis as an outpainting task, the inverse of inpainting, where SD is used to generate everything outside a selected mask using a basic prompt: "generate a clean background." A small improvement in object detection performance was reported for the PASCAL VOC dataset when training data is enhanced with these out-painted images. Lee et al. (2023)[74] apply SD outpainting for military object detection in panoramic smart glasses. They demonstrate that outpainting panoramic backgrounds leads to a 2.5× improvement in detection accuracy over standard cropping and concatenation techniques.

3.3 Image-to-image translation

While full-image generation and inpainting provide methods to synthesize training data from scratch or modify incomplete images, image-to-image translation offers a way to transform existing images into new domains. Two

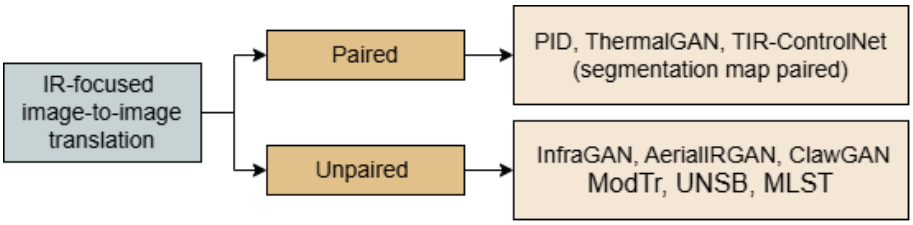

Figure 5: An overview of the models found in literature that focus specifically on RGB-IR domain translation tasks or data augmentation for enhanced performance in downstream AI-tasks on IR imagery.

primary use-cases for image-to-image translation in military applications include environmental and weather adaptations (e.g. day to night, clear to foggy conditions, desert to forest) or modality translations. RGB and IR image data are both important for military scene understanding, yet most available training datasets are RGB-heavy, with limited IR data. Image-to-image translation can be used to convert RGB images into IR representations, expanding synthetic IR training data for tasks such as object detection, classification, and tracking. In our literature scan, we identified a relatively large number of image-to-image translation studies that developed models specifically focused at RGB-to-IR domain translation or data augmentation for enhanced performance in downstream AI-tasks on IR imagery. In addition to the taxonomy in Fig. 1, we therefore present an overview of these IR-focused models in Fig. 5. While translation between other domains is also feasible, there are fewer examples available in the literature. Nonetheless, we present several instances at the end of this section.

Image-to-image translation methods are typically categorized as either paired or unpaired, based on whether they use aligned image pairs for training. Paired methods ensure precise transformations but require one-to-one correspondences, while unpaired methods learn cross-domain mappings without direct supervision.

3.3.1 Paired methods

The pix2pix framework, proposed in 2017,[75] is a widely used approach for supervised image-to-image translation based on conditional GANs (cGANs). Pix2pix requires paired training data, where each input image has a corresponding target image. Although it produces high-quality translations, its reliance on paired data limits its flexibility for military applications. Multiple papers use image-to-image models such as the pix2pix framework[75] (or unpaired models such as CycleGAN,[76] which we will describe in more detail later) to translate RGB images directly to the IR spectrum, extending the models to be more specialized for IR-translation.

ThermalGAN[77] is such a paired image-to-image model, specifically designed for RGB-to-IR transformations, primarily for cross-modality person re-identification of people in IR imagery. The model learns to generate realistic thermal representations of RGB images, by incorporating thermal feature descriptors that capture both the fore- and background characteristics.

More recent papers incorporate diffusion models for RGB-to-IR translation. Mao et al. (2024) introduce the Physics-Informed-Diffusion (PID)[78] model, which integrates physical constraints into a diffusion-based framework to improve IR image fidelity. Unlike previously described generative models, which often treat IR images as a stylistic variation of RGB, PID ensures physical accuracy by applying a TeV decomposition, where T (Temperature), e (emissivity), and V (Thermal Texture) are estimated to model how objects emit and absorb thermal radiation. This physics-informed approach allows the diffusion model to generate realistic IR images that match real-world thermal data better. PID outperforms GAN-based models such as ThermalGAN.[77]

Although the previously discussed studies demonstrate the potential of generative models for RGB-to-IR translation, most rely on datasets with large, easily distinguishable objects, that do not accurately represent military scenarios.

A dataset more representative of military use-case is an UAV dataset titled Aerial Visible-to-Infrared Image Dataset (AVIID) by Han et al. (2023),[79] which provides paired IR and RGB images captured from drones. In their paper, they evaluate both paired and unpaired image-to-image translation methods. Their findings indicate that paired methods, such as pix2pix,[75] generally outperform unpaired methods like CycleGAN.[76]

Although pix2pix was already introduced in 2017, GAN-based methods remain highly relevant for image-to-image translation tasks. More recent advancements continue to refine their capabilities. For example, in a recent study by Parmar et al. (2024),[80] the authors propose pix2pix-Turbo, which is aimed at paired tasks such as sketch-to-image translation. Their model combines single-step diffusion (SD-Turbo,[81] and see Section 3.1.1 for more details on the original SD) with GAN learning objectives, to perform efficient image-to-image translation. Instead of the traditional iterative denoising process in diffusion models (Section 2.3), SD-Turbo compresses the process into a single step, significantly reducing inference time while still leveraging the internal knowledge of the pre-trained diffusion model and thus generating high-quality outputs. LoRA is employed to adapt the original network to new controls and domains, which is described in more detail in Section 3.1.2.[53]

3.3.2 Unpaired methods

While paired image-to-image translation methods are well-suited for precise domain adaptation tasks, their reliance on aligned datasets limits their applicability. The next section describes unpaired techniques, highlighting their flexibility for environmental and weather adaptation as well as modality translation. First, we introduce general-purpose base unpaired methods, then introduce several GAN-based methods focused on RGB-to-IR modality translation. Next, we highlight methods that optimize directly for downstream performance. Finally, we describe methods that integrate diffusion models.

Base models AI models often show a decreased performance in adverse weather conditions or other scenarios they were not specifically trained on.[82–84] While this has been widely explored in the automotive sector, military applications similarly require robustness to changing conditions. Unpaired image-to-image translation has been explored as a solution. One of the most well-known methods is CycleGAN,[76] which learns bidirectional mappings between two domains using two generators and two discriminators. It employs a cycle-consistency loss to ensure that an image translated to the target domain and back retains its original structure. Rothmeier et al. (2021)[82] applied CycleGAN to generate synthetic fog, snow, and rain from clear-weather images. Their study compared CycleGAN against other architectures (e.g., pix2pix,[75] UNIT,[85] and MUNIT[86]), showing that while CycleGAN produced strong results, it struggled with output consistency and controllability.

To address diversity in outputs, Multimodal UNsupervised Image-to-Image Translation (MUNIT)-framework[86] was proposed as a VAE-GAN hybrid that decomposes images into a domain-invariant content space and a domain-specific style space. This disentanglement allows multiple outputs to be generated by combining content from the input with different sampled styles. The authors evaluated MUNIT on a variety of tasks, including seasonal translation and edge-to-photo synthesis. In military contexts, this architecture could facilitate the generation of diverse IR outputs from a single RGB input while preserving spatial structure.

While CycleGAN and MUNIT focus on domain translation, StyleGAN[87] represents a shift toward attribute-controlled synthesis. It introduces a style-based generator architecture that enables highly controllable and realistic image synthesis. Unlike traditional GANs, where the latent vector z is directly fed into the generator (see Section 2.2), StyleGAN first maps z into an intermediate latent space W. This disentanglement step makes it easier to control specific image attributes, allowing attribute-aware image manipulation in W-space. Once trained, a latent vector z can be sampled, mapped to the W-space, and modified to control factors like pose, texture, and lighting. This architecture has been used by Zhang et al. (2022)[88] to create 3D-aware images by manipulating the latent space to generate different viewpoints.

GAN-based RGB-to-IR translation The base models introduced above provide flexible frameworks for general-purpose domain translation. However, many military imaging applications require translating between fundamentally different modalities, most notably from RGB-to-IR, which presents additional challenges, such as structural preservation. To address this, a range of GAN-based methods have been developed specifically for this translation. InfraGAN[89] was designed to generate both RGB and IR images from a given input. It focuses on preserving the structural consistency between modalities using a structural similarity loss function, ensuring that objects retain their original shape and features when converted between RGB and IR.

Building on similar goals, ClawGAN[90] extends the CycleGAN framework[76] by specifically targeting the preservation of structural details in facial image translation between RGB and IR domains. The model introduces

claw-connections for improved feature retention. These are a structural enhancement to the generator network, designed to preserve fine-grained features by combining activations across multiple encoder and decoder layers. While the model supports both RGB-to-IR and IR-to-RGB translations, its primary focus is on IR-to-RGB conversion, aiming to improve facial recognition in low-light conditions where IR imagery is commonly used.

AerialIRGAN[91] introduces an unpaired translation framework that integrates a lightweight CNN with semantic segmentation information from the Segment Anything model (SAM),[92] which helps to preserve object boundaries. The model employs a structural appearance consistency loss, to enforce structural similarity with the RGB image while ensuring IR-specific appearance characteristics. Compared to traditional GAN-based methods (CycleGAN,[76] UNIT,[85] MUNIT[86]) and more recent diffusion-based models (BBDM,[93] CycleGAN-Turbo[80]), AerialIRGAN achieves superior realism on two aerial datasets, including AVIID. However, its evaluation is limited to perceptual metrics to measure the generated image quality, without assessing downstream task performance. Zhan et al. (2024)[94] do extend this (for both CycleGAN[85] and pix2pix[75]) by also incorporating an object detection task on a subset of AVIID, moving beyond only comparing the realism of the generated IR images. Their results demonstrate that assessing task performance (e.g., object detection accuracy) provides a more meaningful evaluation of synthetic IR images compared to image-quality metrics alone. The next methods are explicitly designed with this goal in mind: optimizing for downstream performance.

Downstream task optimization While previous approaches like InfraGAN and ClawGAN aim to translate RGB-to-IR through deep GAN architectures, these methods can introduce visual artifacts or struggle with structure preservation. In contrast, Meta-Learning Style Transfer (MLST)[95] proposes an alternative approach, by optimizing combinations of simple, interpretable filters for the RGB-to-IR task. MLST uses a meta-learning framework, based upon Faster Augment,[96] to learn the best filter compositions that maximize downstream object detection performance. This paper shows that their combined approach outperforms earlier data-driven methods such as InfraGAN[89] and ThermalGAN.[77] However, MLST has not yet been compared to more recent methods such as CycleGAN-Turbo,[80] which also aim to improve structure preservation through hybrid GAN–diffusion architectures.

The work described in the previous paragraphs focuses on RGB-to-IR translation. Medeiros et al. (2024)[97] reverse the problem, and instead focus on learning IR-to-RGB translation. They propose Modality Translator (ModTr), a method that learns a U-Net-based transformation to convert IR images into a RGB-like representation, allowing pre-trained RGB object detectors to process IR data without needing to retrain them. The advantage here is that instead of optimizing for human-perceptible realism, ModTr is directly optimized for the downstream object detection task, improving performance over full IR fine-tuning. While the method does require labeled IR annotations, it eliminates the need for paired RGB-IR data, making it a practical approach for military applications.

Diffusion-based methods While GANs have dominated unpaired translation, recent work has explored diffusion models for this task. While diffusion models have shown remarkable progress in full-image generation and paired image-to-image translation, they have limitations in unpaired image-to-image translation tasks due to the Gaussian prior assumption. In standard diffusion models, image generation begins with a sample from a standard Gaussian distribution, which is iteratively denoised into a target image (Section 2.3). This process relies on the assumption that the data distribution can be smoothly mapped from Gaussian noise, which works well for paired settings with aligned source-target relationships. However, in unpaired image-to-image translation, where source and target domains differ structurally and lack pixel-wise correspondence, this Gaussian prior assumption limits the model's ability to learn meaningful cross-domain mappings. To overcome this, Kim et al. (2024)[98] propose using Schrödinger Bridge, which learns an stochastic differential equations to translate between two arbitrary distributions. Their method, called Unpaired image-to-image translation via Neural Schrodinger Bridge (UNSB), was used to solve various unpaired translation tasks. Vo et al. (2024)[99] combined UNSB with Grounded-SAM[92] to first distinguish between object and background regions, and then translate from RGB-to-IR. Their results show that using only ∼2000 RGB & IR (unpaired) images, it is possible to achieve an acceptable IR ship classification performance. In addition, by using the synthetic IR images, the authors show that ship classification accuracy task can be improved in comparison with models trained only on real TIR images.

Mayr et al. (2024)[100] propose an diffusion-based approach (TIR-ControlNet) to enhance segmentation performance on IR images. Their method retrains ControlNet on 1500 real IR images from the FMB Dataset,[101] conditioning the diffusion process on the segmentation maps. Using this retrained model, they generate over 120,000 diverse synthetic IR images by applying different random seeds to the same segmentation maps. The authors demonstrate that training a SegFormer segmentation model on these synthetic images achieves a performance near to models trained on real data, outperforming previous GAN-based approaches. Their results show that with limited real samples and diffusion-based synthesis, they can effectively narrow the synthetic-to-real gap.

In the previously mentioned study by Parmar et al. (2024), which introduced the paired method pix2pix-Turbo, the authors also propose CycleGAN-Turbo.[80] CycleGAN-Turbo is compared to other unpaired state-of-the-art methods, both GAN-based such as a regular CycleGAN,[76] Contrastive Unpaired Translation (CUT),[102] Instructpix2pix,[103] and diffusion-based such as Stochastic Differential Editing (SDEdit),[104] Plug-and-Play,[105] pix2pix-zero,[106] CycleDiffusion,[107] and Diffusion-Based Image Bridging (DDIB).[108] While CycleGAN-Turbo outperforms CycleDiffusion and DDIB in terms of speed and efficiency, they achieve a comparable performance in terms of realism and domain adaptation. In Fig. 6 we show some visual examples of a CycleGAN-Turbo model fine-tuned for a RGB-to-IR translation task of military vehicle data.

In summary, unpaired image-to-image translation methods offer flexible solutions for generating synthetic data in the absence of aligned datasets. While early GAN-based methods like CycleGAN and MUNIT laid the groundwork for domain translation, more recent work has introduced models specifically adapted to RGB-to-IR conversion, structural preservation, and task optimization. Lightweight techniques such as MLST and purpose-driven models like ModTr demonstrate that performance on downstream tasks can be improved without relying on visually perfect outputs. Meanwhile, diffusion-based approaches—including UNSB, TIR-ControlNet, and CycleGAN-Turbo—show promise in combining realism, structural fidelity, and diversity. Ultimately, the choice of method should be guided by the intended application, whether it prioritizes visual realism, structural accuracy, or direct impact on downstream model performance.

3.3.3 Image-to-image translation for non-EO image data

Image-to-image translation techniques are also used in relation to SAR data. One key application is despeckling, which is the removal of speckle noise, the dominant noise source in SAR imagery. Lattari et al. (2023)[109] proposed CycleSAR based on CycleGAN and more recently Hu et al. (2024)[110] introduced a diffusion-based method, both leading to successful despeckling of SAR images. Another application is the translation of RGB-to-SAR, since SAR images are more difficult to collect and more time-consuming to annotate. Shi et al. (2022)[111] propose a GAN-based method to reduce the appearance discrepancy between the optical and SAR images and showed superior SAR ship detection performance with unlabeled SAR images. Others have focused on the inverse translation of SAR-to-RGB, for example to pseudo-color SAR images to improve interpretability. For example, Shen et al. (2024)[112] introduced baselines for SAR colorization and a translation method called cGAN4ColSAR, based on the pix2pix method, specifically designed for the purpose of colorization. Similar to the work on RGB-to-IR translation, Wang et al. (2024)[113] used a method based on Schrödinger bridge network for SAR-to-RGB image translation and report improved performance over CycleGAN.

Similarly for sonar image data, image-to-image translation techniques have been developed. Sung et al. (2019)[114] propose a method to translate actual sonar images to simulated-like images using a GAN. Liu et al. (2019)[115] use a CycleGAN for realistic image dataset generation for forward-looking sonar, based on camera images. Moreover, Cheng et al. (2022)[116] use RGB images as well as SAR images to generate underwater side-scan sonar images.

4. DISCUSSION

GenAI offers numerous opportunities for synthesizing image data, and recent advances have demonstrated promising results. In this study, we reviewed the opportunities of generating image data for training AI for military applications. We identified three key strategies: full-image generation, inpainting and image-to-image translation. Each strategy presents unique advantages that meet the requirements of different use-cases. A combination of methods is likely required to obtain an optimal result, as can be observed in recent approaches.[117]

Figure 6: Translation of RGB-to-IR images using a CycleGAN-Turbo fine-tuned on an in-house dataset.

Diffusion models represent the current state-of-the-art in full-image generation of RGB data, particularly in text-to-image generation, where training with large datasets yields models with impressive zero-shot performance in generating photorealistic images across diverse domains. However, these models lack specific knowledge in the military domain, making off-the-shelve models insufficient for high-quality full-image generation in military applications. Fine-tuning techniques, such as LoRA,[53] can adapt these models effectively with a relatively limited number of examples. Additionally, optimizing prompts can enhance text-based guidance, while other guidance techniques, such as ControlNet,[42] can further refine image generation by utilizing reference images or depth maps. One challenge of full-image generation is the lack of automatic annotations, e.g. of the object locations, which typically come for free when data is simulated with traditional methods. For non-EO data, most GenAI development focuses on the use of GAN models, although some diffusion-based methods have been proposed.

Diffusion-based inpainting methods produce highly realistic visual outputs, but off-the-shelf models often lack domain-specific knowledge relevant to military contexts. Fine-tuning is likely required for inpainting military-relevant objects, whereas for more generic additions such as vegetation or plastic covers used for concealment, pretrained models may suffice. The primary motivation for using inpainting lies in the abundance of background imagery from drone reconnaissance, general satellite data, and photos available on the internet, while images with appropriate foreground objects are scarce. Inpainting enables the realistic reuse of foreground objects across varied backgrounds. Because inpainting modifies only part of an image, it may require less training data compared to full-image generation. Moreover, similar to ControlNet,[42] inpainting approaches allow greater spatial control over object placement. However, the extent to which inpainting can effectively reduce data requirements remains an open question. Recent state-of-the-art models such as AnyDoor[70] further expand inpainting capabilities by enabling fine-tuning on video data, supporting more advanced harmonization and viewpoint variation based on new masks and backgrounds.

Gen-AI based image-to-image translation techniques show strong potential for adapting images to different environments, such as converting existing datasets to adverse weather conditions or modality transfers. In particular, translation from RGB-to-IR is highly relevant for military applications, where IR data is often scarce but RGB imagery is more readily available. Translating the variation in RGB datasets to IR could substantially enhance the quality of IR training data. It should be noted here that there can be additional variation in the target domain compared to the source domain, such as the relevance of thermal history, which is relevant for IR images, but not visible in RGB images. Capturing this kind of physically meaningful information remains an

open challenge for future research. For translating images to other modalities, such as SAR, promising results have been reported, though the number of studies in this area is still relatively limited. While GAN-based methods like pix2pix[75] and MUNIT[86] are still considered state-of-the-art, combinations with diffusion models, such as CycleGAN-Turbo,[80] are emerging as promising alternatives. Ideally, the development of image-to-image translation models would utilize paired data; however, due to its scarcity, most advancements are made using unpaired data. A key limitation across current methods is generalization. Many image-to-image translation models are trained on relatively small or homogeneous datasets, resulting in degraded performance when when a model is applied to images with a different appearance. Addressing this issue is essential to make image-to-image translation reliably applicable.

Despite the promising visual results of GenAI methods for creating image data, measuring the quality and usefulness of this data for training AI models remains an open challenge. Although a wide range of image quality metrics have been proposed, such as the Frechet Inception Distance (FID) and embedding-based measures, these do not necessarily relate directly to downstream model performance. Even embedding-based metrics can be misleading, as they often rely on CNN backbones trained on RGB data, which may not align with the target domain, such as IR or SAR. For the scope of this review, we have not focused on this issue and do not provide a comparison or evaluation of synthetic image quality assessment techniques. However, we emphasize that establishing standardized and task-relevant evaluation protocols for the generated data is important for future work. Another open question concerns the extent to which GenAI models can meaningfully contribute to domains in which they have seen little or no training data. While prompt engineering can introduce useful variation and even encode prior knowledge, further research is needed to understand how reliably these models extrapolate beyond their training distribution.

GenAI methods carry the risk of hallucinations, which may reduce the performance of subsequently trained AI models. However, the exact impact of these hallucinations is not yet known. Ultimately, the goal is to develop AI methods that perform effectively on real-world data. Evaluating their applicability to military use-cases remains an area for future investigation. If the limitations, as described above, are addressed, GenAI methods show great promise to improve the effectiveness of AI for automated scene understanding in a military context.

REFERENCES

[1] Heslinga, F. G., Ruis, F., Ballan, L., van Leeuwen, M. C., Masini, B., van Woerden, J. E., den Hollander, R. J. M., Berndsen, M., Baan, J., Dijk, J., and Huizinga, W., "Leveraging temporal context in deep learning methodology for small object detection," in [*Artificial Intelligence for Security and Defence Applications*], **12742**, SPIE Sensors + Imaging (2023).

[2] van Leeuwen, M. C., Fokkinga, E. P., Huizinga, W., Baan, J., and Heslinga, F. G., "Toward versatile small object detection with Temporal-YOLOv8," *Sensors* **24**(22) (2024).

[3] Wang, N., Li, X., Suo, Z., Fan, J., Wang, J., and Xie, D., "Traversability analysis and path planning for autonomous wheeled vehicles on rigid terrains," *Drones* **8**(9) (2024).

[4] Dijk, J., Burghouts, G. J., Katyal, K., Yeh, B. Y., Knuth, C., Fokkinga, E., Kasarla, T., and Mettes, P., "Lightweight uncertainty quantification with simplex semantic segmentation for terrain traversability," in [*ICRA 2024 Workshop on Resilient Off-road Autonomy*], (2024).

[5] Ballan, L., Melo, J. G. O., van den Broek, S. P., Baan, J., Heslinga, F. G., Huizinga, W., Dijk, J., and Dilo, A., "EO and radar fusion for fine-grained target classification with a strong few-shot learning baseline," in [*Signal Processing, Sensor/Information Fusion, and Target Recognition XXXIII*], **13057**, SPIE Defense + Commercial Sensing (2024).

[6] Heslinga, F. G., Fokkinga, E. P., Eker, T. H., Liezenga, A. M., den Hollander, R. J. M., Oppeneer, V. O., van Heteren, A. M., van Vossen, R., Kuijf, H. J., van de Sande, J. J. M., van der Burg, D. W., Weyland, L. F., Henderson, H. C., Schadd, M. P. D., and Schutte, K., "On the use of simulated data for target recognition and mission planning," in [*Artificial Intelligence for Security and Defence Applications II*], **13206**, SPIE Sensors + Imaging (2024).

[7] Eker, T. A., Heslinga, F. G., Ballan, L., den Hollander, R. J., and Schutte, K., "The effect of simulation variety on a deep learning-based military vehicle detector," in [*Artificial Intelligence for Security and Defence Applications*], **12742**, 183–196, SPIE Sensors + Imaging (2023).

[8] Heslinga, F. G., Eker, T. A., Fokkinga, E. P., van Woerden, J. E., Ruis, F. A., den Hollander, R. J. M., and Schutte, K., "Combining simulated data, foundation models, and few real samples for training object detectors," in [*Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications II*], **13035**, SPIE Defense + Commercial Sensing (2024).

[9] Ruis, F. A., Liezenga, A. M., Heslinga, F. G., Ballan, L., den Hollander, R. J., van Leeuwen, M. C., Masinia, B., Dijk, J., and Huizinga, W., "Improving object detector training on synthetic data by starting with a strong baseline methodology," in [*Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications II*], **13035**, SPIE Defense + Commercial Sensing (2024).

[10] Fokkinga, E. P., te Hofsté, M. E., den Hollander, R. J., van der Meer, R., Benders, F. P., ter Haar, F. B., Marquis, V. E., van Berkel, M., Voogd, J. M., Eker, T. A., et al., "The validation of simulation for testing deep-learning-based object recognition," in [*Artificial Intelligence for Security and Defence Applications II*], **13206**, 253–275, SPIE (2024).

[11] Moate, C. P., Hayward, S. D., Ellis, J. S., Russell, L., Timmerman, R. O., Lane, R. O., and Strain, T. J., "Vehicle detection in infrared imagery using neural networks with synthetic training data," in [*Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15*], 453–461, Springer (2018).

[12] Westlake, S. T., Volonakis, T. N., Jackman, J., James, D. B., and Sherriff, A., "Deep learning for automatic target recognition with real and synthetic infrared maritime imagery," in [*Artificial intelligence and machine learning in defense applications II*], **11543**, 41–53, SPIE Security + Defence (2020).

[13] Heslinga, F. G., Uysal, F., van Rooij, S. B., Berberich, S., and Caro Cuenca, M., "Few-shot learning for satellite characterisation from synthetic inverse synthetic aperture radar images," *IET Radar, Sonar & Navigations* **18**(4), 649–656 (2024).

[14] van de Sande, J. J. M., Huizinga, W., den Hollander, R. J. M., van den Burg, D. W., and van Vossen, R., "Application of a convolutional neural network trained with simulated low-frequency synthetic aperture sonar data for classification of buried UXO," in [*Proceedings of Underwater Acoustics Conference and Exhibiition (UACE), Greece*], (2023).

[15] Eker, T. A., Fokkinga, E. P., Heslinga, F. G., and Schutte, K., "Balancing 3D-model fidelity for training a vehicle detector on simulated data," in [*Artificial Intelligence for Security and Defence Applications II*], **13206**, SPIE Sensors + Imaging (2024).

[16] Labs, B. F., "FLUX." https://github.com/black-forest-labs/flux (2024).

[17] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I., "Zero-shot text-to-image generation," in [*International conference on machine learning*], 8821–8831, Pmlr (2021).

[18] Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., and Fleet, D. J., "Synthetic data from diffusion models improves ImageNet classification," *ArXiv* **abs/2304.08466** (2023).

[19] Du, X., Sun, Y., Zhu, J., and Li, Y., "Dream the impossible: Outlier imagination with diffusion models," in [*Advances in Neural Information Processing Systems*], **36**, 60878–60901 (2023).

[20] Bansal, G., Nawal, A., Chamola, V., and Herencsar, N., "Revolutionizing visuals: The role of generative ai in modern image generation," *ACM Transactions on Multimedia Computing, Communications and Applications* **20**(11), 1–22 (2024).

[21] Wang, L., Chen, W., Yang, W., Bi, F., and Yu, F. R., "A state-of-the-art review on image synthesis with generative adversarial networks," *IEEE Access* **8**, 63514–63537 (2020).

[22] Alotaibi, A., "Deep generative adversarial networks for image-to-image translation: A review," *Symmetry* **12**(10), 1705 (2020).

[23] Shamsolmoali, P., Zareapoor, M., Granger, E., Zhou, H., Wang, R., Celebi, M. E., and Yang, J., "Image synthesis with adversarial networks: A comprehensive survey and case studies," *Information Fusion* **72**, 126–146 (2021).

[24] Alhabeeb, S. K. and Al-Shargabi, A. A., "Text-to-image synthesis with generative models: Methods, datasets, performance metrics, challenges, and future direction," *IEEE Access* (2024).

[25] Zhan, Z., Chen, D., Mei, J.-P., Zhao, Z., Chen, J., Chen, C., Lyu, S., and Wang, C., "Conditional image synthesis with diffusion models: A survey," *arXiv preprint arXiv:2409.19365* (2024).

[26] Zhang, C., Zhang, C., Zhang, M., and Kweon, I. S., "Text-to-image diffusion models in generative AI: A survey," *arXiv preprint arXiv:2303.07909* (2023).

[27] Baraheem, S. S., Le, T.-N., and Nguyen, T. V., "Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook," *Artificial Intelligence Review* **56**(10), 10813–10865 (2023).

[28] Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A. S., Al-dabbagh, B. S. N., Fadhel, M. A., Manoufali, M., Zhang, J., Al-Timemy, A. H., et al., "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data* **10**(1), 46 (2023).

[29] Kingma, D. P., "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114* (2013).

[30] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial nets," *Advances in neural information processing systems* **27** (2014).

[31] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A., "Generative adversarial networks: An overview," *IEEE signal processing magazine* **35**(1), 53–65 (2018).

[32] McCloskey, B. J., Cox, B. A., Champagne, L., and Bihl, T. J., "Benefits of using blended generative adversarial network images to augment classification model training data sets," *The Journal of Defense Modeling and Simulation* , 15485129231170225 (2023).

[33] Wang, P., Ma, Z., Dong, B., Liu, X., Ding, J., Sun, K., and Chen, Y., "Generative data augmentation by conditional inpainting for multi-class object detection in infrared images," *Pattern Recognition* **153**, 110501 (2024).

[34] Koo, S., Youm, S., and Shin, J., "Cycle-gan-based synthetic sonar image generation for improved underwater classification," in [*Ocean Sensing and Monitoring XVI*], **13061**, 69–83, SPIE Defense + Commercial Sensing (2024).

[35] Zhang, J., Liu, Z., Jiang, W., Liu, Y., Zhou, X., and Li, X., "Application of deep generative networks for SAR/ISAR: a review," *Artificial Intelligence Review* **56**(10), 11905–11983 (2023).

[36] Ulhaq, A. and Akhtar, N., "Efficient diffusion models for vision: A survey," *arXiv preprint arXiv:2210.09292* (2022).

[37] Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M., "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(9), 10850–10869 (2023).

[38] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B., "High-resolution image synthesis with latent diffusion models," (2021).

[39] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M., "Photorealistic text-to-image diffusion models with deep language understanding," (2022).

[40] Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A., "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114* (2021).

[41] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al., "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in neural information processing systems* **35**, 25278–25294 (2022).

[42] Zhang, L., Rao, A., and Agrawala, M., "Adding conditional control to text-to-image diffusion models," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 3836–3847 (2023).

[43] Esser, P., Rombach, R., and Ommer, B., "Taming transformers for high-resolution image synthesis," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 12873–12883 (2021).

[44] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R., "SDXL: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952* (2023).

[45] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I., "Learning transferable visual models from natural language supervision," *CoRR* **abs/2103.00020** (2021).

[46] Peebles, W. and Xie, S., "Scalable diffusion models with transformers," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 4195–4205 (2023).

[47] Liu, H., Li, C., Wu, Q., and Lee, Y. J., "Visual instruction tuning," *Advances in neural information processing systems* **36**, 34892–34916 (2023).

[48] Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al., "Scaling rectified flow transformers for high-resolution image synthesis," in [*Forty-first international conference on machine learning*], (2024).

[49] Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., et al., "SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers," in [*The Thirteenth International Conference on Learning Representations*], (2025).

[50] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D., "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618* (2022).

[51] Ye, H., Zhang, J., Liu, S., Han, X., and Yang, W., "IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721* (2023).

[52] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K., "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 22500–22510 (2023).

[53] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W., "LoRA: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685* (2021).

[54] Zhang, B., Zhang, P., Dong, X., Zang, Y., and Wang, J., "Long-CLIP: Unlocking the long-text capability of CLIP," in [*European Conference on Computer Vision*], 310–325, Springer (2024).

[55] Dhariwal, P. and Nichol, A., "Diffusion models beat gans on image synthesis," in [*Proceedings of the 35th International Conference on Neural Information Processing Systems*], *NIPS '21*, Curran Associates Inc., Red Hook, NY, USA (2021).

[56] Ho, J. and Salimans, T., "Classifier-free diffusion guidance," (2022).

[57] Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., and Li, H., "Semantic image synthesis via diffusion models," *arXiv preprint arXiv:2207.00050* (2022).

[58] Rothmeier, T., Huber, W., and Knoll, A. C., "Time to shine: Fine-tuning object detection models with synthetic adverse weather images," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], 4447–4456 (2024).

[59] Huang, Z., Zhang, X., Tang, Z., Xu, F., Datcu, M., and Han, J., "Generative artificial intelligence meets synthetic aperture radar: A survey," *IEEE Geoscience and Remote Sensing Magazine* , 2–44 (2025).

[60] Song, Q., Xu, F., Zhu, X. X., and Jin, Y.-Q., "Learning to generate SAR images with adversarial autoencoder," *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–15 (2022).

[61] Ju, M., Niu, B., and Zhang, J., "SAR image generation method for oriented ship detection via generative adversarial networks," *SIViP* **18**, 589–596 (2024).

[62] Zhang, F., Hou, X., Wang, Z., Cheng, C., and Tan, T., "Side-scan sonar image generator based on diffusion models for autonomous underwater vehicles," *Journal of Marine Science and Engineering* **12**(8) (2024).

[63] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M., "GLIDE: towards photorealistic image generation and editing with text-guided diffusion models," *CoRR* **abs/2112.10741** (2021).

[64] Nicola, F., Gennaro, V., and Giovanna, C., "I dream my painting: Connecting MLLMs and diffusion models via prompt generation for text-guided multi-mask inpainting," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], (2025).

[65] Liu, H., Yang, H., Huijben, E. M. C., Schuiveling, M., Su, R., Pluim, J. P. W., and Veta, M., "PathoPainter: Augmenting histopathology segmentation via tumor-aware inpainting," *arXiv preprint arXiv:2503.04634* (2025).

[66] Pérez-García, F., Bond-Taylor, S., Sanchez, P. P., van Breugel, B., Castro, D. C., Sharma, H., Salvatelli, V., Wetscherek, M. T. A., Richardson, H., Lungren, M. P., Nori, A., Alvarez-Valle, J., Oktay, O., and Ilse, M., "RadEdit: Stress-testing biomedical vision models via diffusion image editing," in [*Computer Vision – ECCV 2024*], 358–376, Springer Nature Switzerland, Cham (2024).

[67] Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., and Wen, F., "Paint by example: Exemplar-based image editing with diffusion models," *arXiv preprint arXiv:2211.13227* (2022).

[68] Kim, K., Park, S., Lee, J., and Choo, J., "Reference-based image composition with sketch via structure-aware diffusion model," *arXiv preprint arXiv:2304.09748* (2023).

[69] Zhang, X., Guo, J., Yoo, P., Matsuo, Y., and Iwasawa, Y., "Paste, inpaint and harmonize via denoising: Subject-driven image editing with pre-trained diffusion model," *arXiv preprint arXiv:2306.07596* (2023).

[70] Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., and Zhao, H., "AnyDoor: Zero-shot object-level image customization," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 6593–6602 (2024).

[71] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P., "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193* (2024).

[72] Pichler, A. and Hueber, N., "Method for training deep neural networks in vehicle detection using drone-captured data and background synthesis," in [*Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications II*], **13035**, 150–162, SPIE Defense + Commercial Sensing (2024).

[73] Li, Y., Dong, X., Chen, C., Zhuang, W., and Lyu, L., "A simple background augmentation method for object detection with diffusion model," *arXiv preprint arXiv:2408.00350* (2024).

[74] Lee, J., Moon, S., and Nam, D., "A study on military object detection in panoramic view using stable diffusion," in [*2023 14th International Conference on Information and Communication Technology Convergence (ICTC)*], 1889–1891 (2023).

[75] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A., "Image-to-image translation with conditional adversarial networks," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 1125–1134 (2017).

[76] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in [*Proceedings of the IEEE international conference on computer vision*], 2223–2232 (2017).

[77] Kniaz, V. V., Knyaz, V. A., Hladůvka, J., Kropatsch, W. G., and Mizginov, V. A., "ThermalGAN: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset," in [*Computer Vision – ECCV 2018 Workshops*], Springer International Publishing (2018).

[78] Mao, F., Mei, J., Lu, S., Liu, F., Chen, L., Zhao, F., and Hu, Y., "PID: Physics-informed diffusion model for infrared image generation," *arXiv preprint arXiv:2407.09299* (2024).

[79] Han, Z., Zhang, Z., Zhang, S., Zhang, G., and Mei, S., "Aerial visible-to-infrared image translation: Dataset, evaluation, and baseline," *Journal of remote sensing* **3**, 0096 (2023).

[80] Parmar, G., Park, T., Narasimhan, S., and Zhu, J.-Y., "One-step image translation with text-to-image models," *arXiv preprint arXiv:2403.12036* (2024).

[81] Sauer, A., Lorenz, D., Blattmann, A., and Rombach, R., "Adversarial diffusion distillation," in [*European Conference on Computer Vision*], 87–103, Springer (2024).

[82] Rothmeier, T. and Huber, W., "Let it snow: On the synthesis of adverse weather image data," *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC* , 3300–3306 (2021).

[83] Rothmeier, T., Wachtel, D., Von Dem Bussche-Hunnefeld, T., and Huber, W., "I had a bad day: Challenges of object detection in bad visibility conditions," *IEEE Intelligent Vehicles Symposium* (2023).

[84] Hasirlioglu, S. and Riener, A., "Challenges in object detection under rainy weather conditions," in [*Intelligent Transport Systems, From Research and Development to the Market Uptake: Second EAI International Conference, INTSYS 2018, Guimarães, Portugal, November 21–23, 2018, Proceedings 2*], 53–65, Springer (2019).

[85] Liu, M.-Y., Breuel, T., and Kautz, J., "Unsupervised image-to-image translation networks," *Advances in neural information processing systems* **30** (2017).

[86] Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J., "Multimodal unsupervised image-to-image translation," in [*Proceedings of the European conference on computer vision (ECCV)*], 172–189 (2018).

[87] Karras, T., "A style-based generator architecture for generative adversarial networks," *arXiv preprint arXiv:1812.04948* (2019).

[88] Zhang, Y., Chen, W., Ling, H., Gao, J., Zhang, Y., Torralba, A., and Fidler, S., "Image GANs meet differentiable rendering for inverse graphics and interpretable 3D neural rendering," *CoRR* **abs/2010.09125** (2020).

[89] Özkanoğlu, M. A. and Ozer, S., "InfraGAN: A GAN architecture to transfer visible images to infrared domain," *Pattern Recognition Letters* **155**, 69–76 (2022).

[90] Luo, Y., Pi, D., Pan, Y., Xie, L., Yu, W., and Liu, Y., "ClawGAN: Claw connection-based generative adversarial networks for facial image translation in thermal to RGB visible light," *Expert Systems with Applications* **191**, 116269 (2022).

[91] Ma, D., Su, J., Li, S., and Xian, Y., "AerialIRGAN: unpaired aerial visible-to-infrared image translation with dual-encoder structure," *Scientific Reports* **14**(1), 22105 (2024).

[92] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al., "Segment anything," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 4015–4026 (2023).

[93] Li, B., Xue, K., Liu, B., and Lai, Y.-K., "BBDM: Image-to-image translation with brownian bridge diffusion models," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*], 1952–1961 (2023).

[94] Zhan, X., Cui, S., Feng, G., Zang, Q., Chen, Z., Liu, W., Lv, Y., and Gan, H., "Evaluation of the generated visible-to-infrared images using cyclegan and pix2pix based on pre-sorted aviid-3 dataset," in [*Conference on Spectral Technology and Applications (CSTA 2024)*], **13283**, 946–954, SPIE (2024).

[95] Stump, E. A., Luzi, F., Collins, L. M., and Malof, J. M., "Meta-learning for color-to-infrared cross-modal style transfer," *arXiv preprint arXiv:2212.12824* (2022).

[96] Hataya, R., Zdenek, J., Yoshizoe, K., and Nakayama, H., "Faster autoaugment: Learning augmentation strategies using backpropagation," in [*Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*], 1–16, Springer (2020).

[97] Medeiros, H. R., Aminbeidokhti, M., Peña, F. A. G., Latortue, D., Granger, E., and Pedersoli, M., "Modality translation for object detection adaptation without forgetting prior knowledge," in [*European Conference on Computer Vision*], 51–68, Springer (2024).

[98] Kim, B., Kwon, G., Kim, K., and Ye, J. C., "Unpaired image-to-image translation via neural schrödinger bridge," (2024).

[99] Vo, D. T., Duc, P. A., Thao, N. N., and Ninh, H., "An approach to synthesize thermal infrared ship images," in [*Synthetic Data for Computer Vision Workshop@ CVPR 2024*], (2024).

[100] Mayr, C., Kubler, C., Haala, N., and Teutsch, M., "Narrowing the synthetic-to-real gap for thermal infrared semantic image segmentation using diffusion-based conditional image synthesis," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 3131–3141 (2024).

[101] Liu, J., Liu, Z., Wu, G., Ma, L., Liu, R., Zhong, W., Luo, Z., and Fan, X., "Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 8115–8124 (2023).

[102] Park, T., Efros, A. A., Zhang, R., and Zhu, J.-Y., "Contrastive learning for unpaired image-to-image translation," in [*Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*], 319–345, Springer (2020).

[103] Brooks, T., Holynski, A., and Efros, A. A., "Instructpix2pix: Learning to follow image editing instructions," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 18392–18402 (2023).

[104] Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S., "SDEdit: Guided image synthesis and editing with stochastic differential equations," *arXiv preprint arXiv:2108.01073* (2021).

[105] Tumanyan, N., Geyer, M., Bagon, S., and Dekel, T., "Plug-and-play diffusion features for text-driven image-to-image translation," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 1921–1930 (2023).

[106] Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., and Zhu, J.-Y., "Zero-shot image-to-image translation," in [*ACM SIGGRAPH 2023 Conference Proceedings*], 1–11 (2023).

[107] Wu, C. H. and De la Torre, F., "A latent space of stochastic diffusion models for zero-shot image editing and guidance," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 7378–7387 (2023).

[108] Su, X., Song, J., Meng, C., and Ermon, S., "Dual diffusion implicit bridges for image-to-image translation," *arXiv preprint arXiv:2203.08382* (2022).

[109] Lattari, F., Santomarco, V., Santambrogio, R., Rucci, A., and Matteucci, M., "CycleSAR: SAR image despeckling as unpaired image-to-image translation," in [*2023 International Joint Conference on Neural Networks (IJCNN)*], 1–8 (2023).

[110] Hu, X., Xu, Z., Chen, Z., Feng, Z., Zhu, M., and Stanković, L., "SAR despeckling via regional denoising diffusion probabilistic model," in [*IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*], 7226–7230, IEEE (2024).

[111] Shi, Y., Du, L., Guo, Y., and Du, Y., "Unsupervised domain adaptation based on progressive transfer for ship detection: From optical to SAR images," *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–17 (2022).

[112] Shen, K., Vivone, G., Yang, X., Lolli, S., and Schmitt, M., "A benchmarking protocol for SAR colorization: From regression to deep learning approaches," *Neural Networks* **169**, 698–712 (2024).

[113] Wang, J., Yang, H., He, Y., Zheng, F., Liu, Z., and Chen, H., "An unpaired SAR-to-optical image translation method based on Schrödinger bridge network and multi-scale feature fusion," *Scientific Reports* **14**(1), 27047 (2024).

[114] Sung, M., Cho, H., Kim, J., and Yu, S.-C., "Sonar image translation using generative adversarial network for underwater object recognition," in [*2019 IEEE Underwater Technology (UT)*], 1–6 (2019).

[115] Liu, D., Wang, Y., Ji, Y., Tsuchiya, H., Yamashita, A., and and, H. A., "CycleGAN-based realistic image dataset generation for forward-looking sonar," *Advanced Robotics* **35**(3-4), 242–254 (2021).

[116] Cheng, Z., Huo, G., and Li, H., "A multi-domain collaborative transfer learning method with multi-scale repeated attention mechanism for underwater side-scan sonar image classification," *Remote Sensing* **14**(2) (2022).

[117] Zhu, J., Li, S., Liu, Y., Huang, P., Shan, J., Ma, H., and Yuan, J., "ODGEN: Domain-specific object detection data generation with diffusion models," *arXiv preprint arXiv:2405.15199* (2024).