Journal Title
XX(X):1–12

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Diffusion-Transformer Hybrid Framework for Dataset Construction in Deep Learning-Based TEL Detection from Satellite Imagery

Soomin Kwon[1], Jeongin Lee[1], Seongil Jo[1], and Jaeoh Kim[1]

Abstract

Automated identification of transporter-erector-launchers (TELs) in satellite imagery is a critical task for strategic surveillance, yet it remains hindered by extreme data sparsity and the operational complexity of camouflaged military assets. To overcome these challenges, this study proposes a three-stage synthetic data generation framework that integrates semantic-preserving diffusion-based augmentation, context-aware object insertion, and transformer-based image inpainting. In the first stage, TEL instances are augmented using a conditional diffusion model guided by image inputs and textual inversion, enabling variation in camouflage patterns while maintaining structural fidelity. In the second stage, the augmented TELs are inserted into contextually realistic satellite backgrounds—such as tunnel entrances, mountainous regions, and airfield aprons—identified through open-source intelligence. In the third stage, a transformer-based inpainting model is applied to eliminate boundary artifacts and ensure spatial and spectral continuity around the inserted objects. Experimental results demonstrate that the proposed framework outperforms conventional GAN-based approaches, achieving a 38% reduction in Fréchet Inception Distance and producing visually coherent datasets suitable for training deep learning–based detection models. Overall, this study presents a scalable and domain-adaptive approach to synthetic dataset construction, supporting the advancement of defense AI capabilities vital to the transition toward a science-and-technology–oriented military force.

Keywords

Diffusion Model, Satellite Imagery, Synthetic Dataset Generation, transporter-erector-launcher, Transformer Model

1 Introduction

The current international security environment, including the Korean Peninsula, is shaped by multiple regional conflicts that unfold simultaneously and unpredictably. Since the early 2020s, North Korea has revealed solid-fuel intercontinental ballistic missiles and conducted hypersonic tests, while the war in Ukraine has demonstrated the operational impact of dispersed, highly mobile missile forces. In these circumstances, traditional surveillance and reconnaissance systems alone are insufficient for real-time situational awareness and timely decision making. Although super-resolution electro-optical (EO) imagery can be collected and disseminated almost in real time, locating threat targets in such large volumes of data exceeds the capacity of human imagery analysts. In these cases, Deep-learning-based object-detection techniques are therefore viewed as a practical solution and have become a core element of battlefield information-processing pipelines in several countries.

Transporter-Erector-Launchers (TELs) are key strategic assets, as their high mobility and concealment make them exceptionally difficult to detect. Failing to identify them in advance sharply reduces pre-launch reaction time; consequently, the development of image-based artificial intelligence (AI) techniques for early detection is essential. In response to this need, previous research has highlighted the importance of detecting military vehicles and tanks in SAR images and proposed a deep learning-based approach for target detection[1]. However, applying this approach to the task of TEL detection presents significant challenges. In practice, instances of TELs in satellite and SAR imagery are even rarer than those of general military vehicles and tanks, leading to an absolute scarcity of data for training and validation.

Recent target detection studies have tended to modify established architectures such as CNN, YOLO, and DETR to gain higher accuracy[2–4]. However, practical performance depends more on how realistic, large-scale, and diverse the training data are than on the model itself[5,6]. Discussion of TEL-specific dataset construction is relatively limited, and this gap can constrain how far model improvements can go. The main data-related and domain-related challenges are as follows. First, raw data are extremely scarce: Satellite imagery TEL instances are rarely reported due to security restrictions and observation

[1]Department of Statistics and Data Science, Inha University, South Korea

Corresponding author:
Jaeoh Kim Department of Statistics and Data Science, Inha University, Michuhol-gu, Incheon, 22212, South Korea.
Seongil Jo Department of Statistics and Data Science, Inha University, Michuhol-gu, Incheon, 22212, South Korea.
Email: bstatsjo@gmail.com, jaeoh.k@inha.ac.kr

difficulty, and the few open images available are insufficient for training and validation. Consequently, data augmentation is effectively a prerequisite. However, systematic schemes for augmenting TEL objects or generating composite TEL-in-context imagery remain rare and context-preserving TEL augmentation is still uncommon[7]. Second, the structural uniqueness of TELs complicates the augmentation process. A realistic dataset must preserve TEL-specific geometry and camouflage in a true top-down view; generic vehicle augmentation or simple copy-paste techniques do not meet this requirement. Third, the background of the satellite image is also special. TELs are often concealed near tunnel entrances, mountain roads, or airfield aprons, locations dictated by military considerations. Randomly pasting TELs into arbitrary scenes introduces context errors that reduce training effectiveness. Background selection and precise compositing must therefore reflect terrain features and exposure risk.

To address these issues, this study proposes a 3-stage process for synthetic data generation to expand the TEL training data. Each stage corresponds to the challenges of data scarcity, object specificity, and background context.

Stage 1. TEL object augmentation Because only a few original TEL photographs are available, we first augment TEL objects using an Image-to-Image diffusion model combined with textual inversion that encodes model-specific features. The diffusion process varies camouflage color, pattern, and fine structural details while enforcing semantic invariance—that is, the TEL's overall geometry, structure, and critical components remain unchanged.

Stage 2. Context-based object insertion The augmented TEL images are then inserted into real satellite scenes. Scenes are selected from open sources covering Sunan, Geumsu Mountain, Haeju, and Rason, well-known North Korean military hubs. Within each area, plausible TEL positions—tunnel entrances, concealed roads, and runway edges—are chosen based on historical information and military domain knowledge. TELs are composited into the scenes at the appropriate resolution, preserving the realism of the final imagery.

Stage 3. Boundary-consistency inpainting Simple compositing introduces hard edges and pixel discontinuities. A transformer-based inpainting model is applied to reconstruct texture and shading around the object, ensuring continuity so that the detector learns true TEL features rather than splice artifacts.

The contributions of this study are as follows. First, we introduce a semantic-invariant augmentation method that diversifies the appearance of the TEL while preserving its distinctive structure. Second, we design an insertion procedure that selects realistic deployment sites, such as tunnel entrances, mountainous concealment areas, and missile bases near airports, based on open satellite imagery and military activity reports. Third, we adopt a transformer-based inpainting step to reduce boundary artifacts and ensure spectral consistency. Finally, we apply the complete process to real satellite image tiles and a limited set of open-source TEL photographs, thereby constructing a large training dataset and validating its effectiveness through both quantitative and qualitative experiments.

This paper is organized into five sections. Section 2 reviews related works relevant to the proposed techniques. Section 3 details the 3-stage process for synthetic-data generation for TEL detection. It consists of three subsections that cover augmentation, insertion, and inpainting. Section 4 presents simulation results for each subsection. Finally, Section 5 summarizes the contributions, discusses the remaining challenges, and outlines future research directions.

2 Related Works

2.1 Data Augmentation Using Generative Models

Remote sensing imagery, including satellite and aerial images, often suffers from limited data availability due to high acquisition costs and security restrictions. This scarcity poses a major challenge in the training of deep learning models for tasks such as object detection and classification. To alleviate this issue, generative models—most notably Generative Adversarial Networks (GANs)—have been explored as a data augmentation strategy, offering both increased visual diversity and the ability to synthesize images with domain-relevant features and structural consistency.

Wasserstein GAN (WGAN), introduced by Arjovsky et al.[8], addresses the instability and mode collapse in traditional GANs by employing a loss function based on the Wasserstein distance. This allows for more stable discriminator training through the imposition of Lipschitz continuity. Building on this, Karras et al.[9] proposed StyleGAN, which adopts a style transfer framework that allows fine-grained control over generated content, from low-level textures to high-level semantic structures. StyleGAN has been shown to generate high-resolution images with semantic coherence, making it suitable for data synthesis in structurally complex domains. Despite these advantages, GAN-based models exhibit several limitations when applied to operational satellite imagery, particularly for applications involving defense-related imagery. Targets such as TELs require precise shape fidelity and contextual integration. However, GAN-generated synthetic images can exhibit structural deformation, unrealistic object-background transitions, and semantic inconsistencies, factors that can degrade detection performance and reduce the operational reliability of AI-driven analysis systems.

To address the limitations of GANs, recent advances have shifted toward diffusion-based generative models. Ho et al.[10] proposed the Denoising Diffusion Probabilistic Model (DDPM), which synthesizes high-quality images by iteratively reversing a noise injection process. This method provides more stable training and supports better diversity in generated outputs compared to adversarial training. Building upon this, Rombach et al.[11] introduced the Latent Diffusion Model (LDM), which performs the diffusion process within a compressed latent space. This significantly reduces computational requirements while preserving output resolution and quality. Moreover, LDMs are particularly effective for text-to-image generation, allowing users to control image characteristics such as object shape, texture, and viewpoint via textual prompts—an essential feature for domain-specific augmentation.

To extend diffusion models for specialized military applications, Gal et al.[12] proposed Textual Inversion, which enables the training of custom token embeddings corresponding to previously unseen entities (e.g., SA-22 or TEL-A). This allows users to synthesize imagery of domain-specific objects using natural language prompts without changes to the model architecture. In the context of remote sensing, Trabucco et al.[13] demonstrated that diffusion-based augmentation leads to measurable improvements in detection and segmentation tasks, especially under class-imbalanced or low-resource conditions. These results suggest that diffusion models contribute to enhanced performance in downstream tasks by generating data that is visually consistent with target domains and enhancing model robustness across varying conditions.

2.2 Object Insertion and Masking in Satellite Image

In settings where annotated data is limited and object classes are rare or sensitive is often the case in military and remote sensing domains—instance-level augmentation strategies have emerged as a practical solution. Copy-Paste augmentation, in particular, provides a computationally efficient method for dataset expansion with proven performance benefits. By segmenting foreground objects from source imagery and compositing them into novel backgrounds, it facilitates the creation of additional training samples with relatively low resource requirements, thus reducing the need for data collection and manual annotation. Ghiasi et al.[14] demonstrated that Copy-Paste augmentation can significantly improve instance segmentation performance, especially for underrepresented classes. The approach requires neither major modifications to the model architecture nor substantial computational resources, making it an attractive option for scaling datasets. However, its application to high-fidelity domains such as satellite imagery presents domain-specific challenges that must be addressed for effective deployment.

The first limitation involves contextual mismatch: pasted objects often have lighting, resolution, or orientation mismatches with the background, leading to artifacts and perceptual inconsistencies. Second, insertion locations are typically selected at random, without considering the strategic and spatial factors that typically influence object placement in actual satellite imagery such as concealment, line-of-sight constraints, and proximity to infrastructure. Third, ignoring existing background elements during compositing may cause overlaps or unrealistic scaling, which may alter the statistical properties of the training data distribution. Fourth, conventional Copy-Paste methods operate strictly in two dimensions and do not account for depth, terrain morphology, or lighting context, all of which are important considerations in remote sensing applications. To address these limitations, domain-specific constraints must be incorporated into the augmentation pipeline. In military reconnaissance imagery, for example, it is essential to guide object insertion using strategically informed placement rules, context-aware blending techniques, and ideally, physically based rendering (PBR) engines to better emulate the physical and operational realities captured in satellite data.

Another important yet often underexplored aspect of object insertion is the precision of object masking. In satellite imagery, generating accurate masks is challenging due to low signal-to-noise ratios, occlusions, and subtle spectral variations. Conventional binary masking—commonly using solid black or white cutouts—tends to produce sharp, unrealistic boundaries that disrupt visual continuity and fail to reflect the contextual and spectral subtleties of the scene. As a result, inserted objects may appear conspicuously artificial, potentially distorting data distribution and impairing model generalization. To address these issues, recent studies propose more perceptually aligned masking strategies. Zhu et al.[15] highlight the limitations of binary masks in GAN-based object editing and advocate for visually coherent alternatives. Surveys on deep inpainting further underscore the importance of structural fidelity and semantic context in remote sensing[16]. Tools like the Segment Anything Model (SAM) offer promising avenues for generating high-quality masks across diverse imaging conditions. Integrating such refined techniques is essential for producing semantically consistent and physically realistic augmentations in satellite data.

2.3 Inpainting and Restoration in Remote Sensing

Inpainting is a fundamental technique for restoring missing or corrupted image regions using information from the surrounding context. In the domain of satellite imagery, this method is widely used in tasks such as gap-filling, artifact correction, and occlusion removal. Traditional inpainting techniques primarily rely on exemplar-based strategies, which synthesize the missing content by duplicating textures from neighboring patches. For example, the method proposed by Criminisi et al.[17] determines the patch-filling sequence based on similarity metrics and uses surrounding pixel data to fill occluded regions. Although the methods yield visually plausible results, they may exhibit limitations in accurately preserving geometric structures and maintaining semantic continuity.

An alternative approach by Telea employs the Fast Marching Method to propagate known information into missing areas. While computationally efficient, its effectiveness declines in handling complex boundaries and nonhomogeneous patterns—common characteristics of high-resolution satellite imagery with varied terrain and infrastructure. With the advent of deep learning, CNN-based inpainting models have significantly improved the reconstruction of semantically meaningful structures. For example, Yu et al.[18] introduced a contextual attention mechanism to capture non-local dependencies, improving coherence in restored regions. Nazeri et al.[19] further enhanced structure-aware inpainting with EdgeConnect, a two-stage pipeline incorporating edge prediction to guide the fill-in process. Despite these advancements, CNN-based methods face key challenges in remote sensing contexts. Low contrast and ambiguous textures complicate foreground-background separation, while large occlusions often result in blurred or inconsistent reconstructions. Moreover, CNNs struggle to capture long-range

spatial dependencies, limiting their effectiveness in restoring extended features such as road networks or rivers.

Given that satellite imagery frequently spans several kilometers and includes repeating structural elements across distant areas, inpainting models must incorporate mechanisms to capture and leverage long-range spatial dependencies. Transformer-based models, which utilize self-attention to model global interactions, are particularly making them ideal for this task. Unlike CNNs that rely on localized filters, transformers can integrate both local and global contextual information, making them well suited for applications requiring spatial consistency across large image extents. Recent work by Zeng et al.[20] introduced the Mask-Aware Transformer (MAT), a model specifically designed for inpainting tasks involving large missing regions. MAT incorporates the binary mask directly into the attention mechanism, allowing the network to dynamically prioritize relevant contextual cues during the reconstruction process. The approach has shown improved results in structure preservation and visual fidelity relative to CNN-based approaches, based on experimental evaluations.

In conclusion, while traditional and CNN-based methods offer a foundation for image inpainting, the adoption of transformer-based architectures represents a notable advancement in addressing the specific challenges of satellite imagery. These models provide enhanced capabilities for maintaining structural and semantic coherence in large-scale images, making them a promising solution for future inpainting tasks within remote sensing pipelines.

2.4 Summary

This section has examined recent progress in generative techniques for satellite image augmentation and restoration, spanning GAN- and diffusion-based synthesis, as well as inpainting and copy-paste strategies. While these methods have significantly enriched the methodological repertoire for satellite imagery analysis, their application to military reconnaissance remains constrained by several domain-specific limitations. The absence of publicly available datasets specifically annotated for TELs and similar military assets hinders model generalization and restricts the development of task-specific augmentation strategies. Furthermore, common generative models often introduce structural distortions that compromise object fidelity and spatial coherence, potentially degrading detection performance in mission-critical or defense-related applications. In addition, object-insertion methods often do not incorporate intrinsic strategic and operational constraints in military environments, such as concealment tactics, terrain compatibility, and realistic deployment scenarios. Likewise, most inpainting techniques are optimized for urban or natural imagery and are not designed to accommodate the spectral and topographic variability of satellite data, particularly in rural or undeveloped regions. Finally, structure-blind restoration can cause internal TEL features to blend into their surroundings, leading to ambiguity that can affect detection reliability. These observations underscore the need for domain-specific augmentation frameworks that integrate contextual priors, structural consistency, and military-relevant semantics, an approach explored in detail in the following section.

3 Methods

This section provides a detailed methodology for constructing a TEL detection dataset based on satellite images. In practice, as explained in the Backgrounds section, it is often the case that no TEL appears within the satellite imagery. Consequently, it becomes essential to generate training data by directly inserting separately obtained TEL images into satellite images. First, we perform various augmentations on top-down TEL images, as detailed in *Augmentation of TEL Dataset*. Next, TEL objects are masked and inserted into satellite images at specific locations and scales to simulate real-world conditions described in *TEL Object Masking and Insertion*. Finally, we apply a mask restoration procedure, described in *Restoration of Masked Areas in Satellite Images*, to smooth unnatural boundaries and enhance both the visual and structural integrity of the images. As illustrated in Figure 1, this systematic process results in a robust TEL dataset for effective model training. The following sections discuss each component of the methodology in detail.

3.1 Augmentation of TEL Dataset

It is challenging to obtain large quantities of real TEL images. In particular, TEL images for each model are limited in number, and they also vary in resolution, camera angle, and other parameters, making dataset construction difficult. To address these data scarcity issues and ultimately enhance the reliability of deep learning–based TEL detection models, it is crucial to generate synthetic data through augmentation. In the process of creating synthetic data, we must consider the highly specialized and domain-specific requirements of the military and satellite imaging domains. Since TEL objects are to be inserted into satellite images, only top-down perspectives of TELs should be generated, which differs from common object augmentation. Moreover, TELs in a military context typically feature camouflage and monotone coloring. Therefore, a specialized augmentation approach is needed, one that maintains the intrinsic attributes of TELs while reflecting these domain-specific considerations.

The specific objectives of this process are as follows. First, TEL object images must be acquired in a top-down view to replicate the perspective of actual satellite imagery. Second, semantic invariance-preserving the intrinsic semantic features of the objects while allowing for variation—must be ensured. In this context, the fundamental characteristics of TELs, particularly their vehicle configuration and missile launch mechanism, should be maintained, and each model's distinctive attributes, such as size, design, and structural features, should remain uncompromised during augmentation. Third, the augmented data should exhibit sufficient diversity. While maintaining semantic invariance, broad variations in color, camouflage patterns, brightness, and similar attributes must be introduced to ensure that the detection model performs robustly under diverse conditions. Lastly, domain-specific features need to be incorporated, as military camouflage patterns and defense-specific textures differ significantly from those used in typical object augmentation. Consequently, appropriate text prompts and model configurations tailored to the military and satellite domains must be adopted.

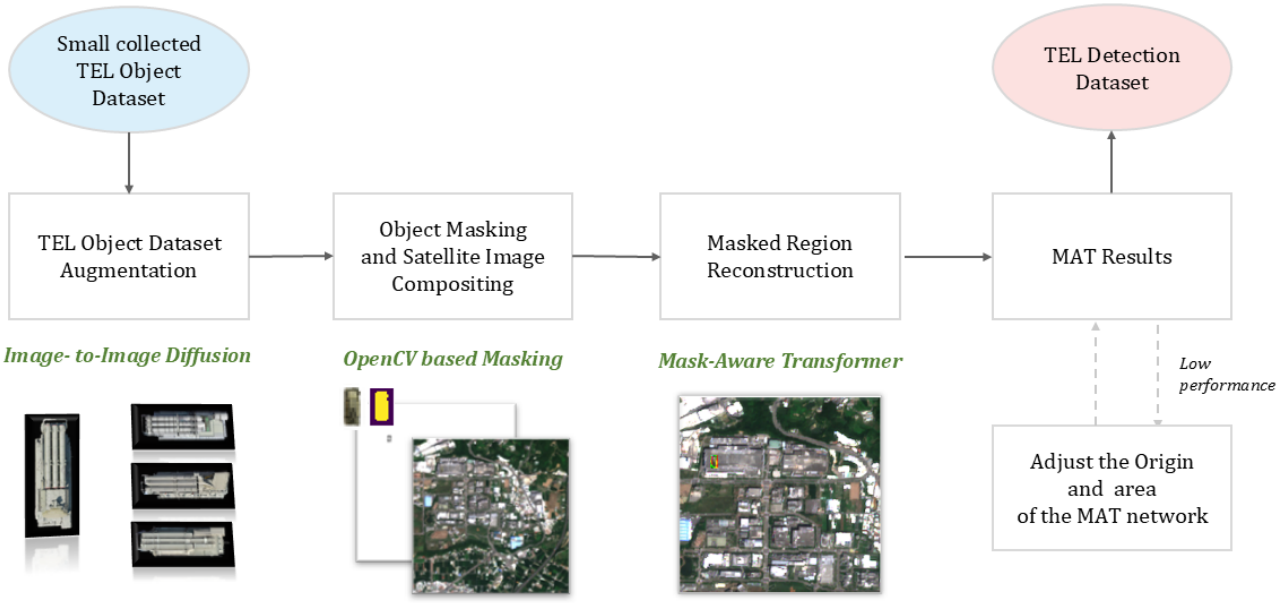

Figure 1. Overview of Generation for TEL detection dataset

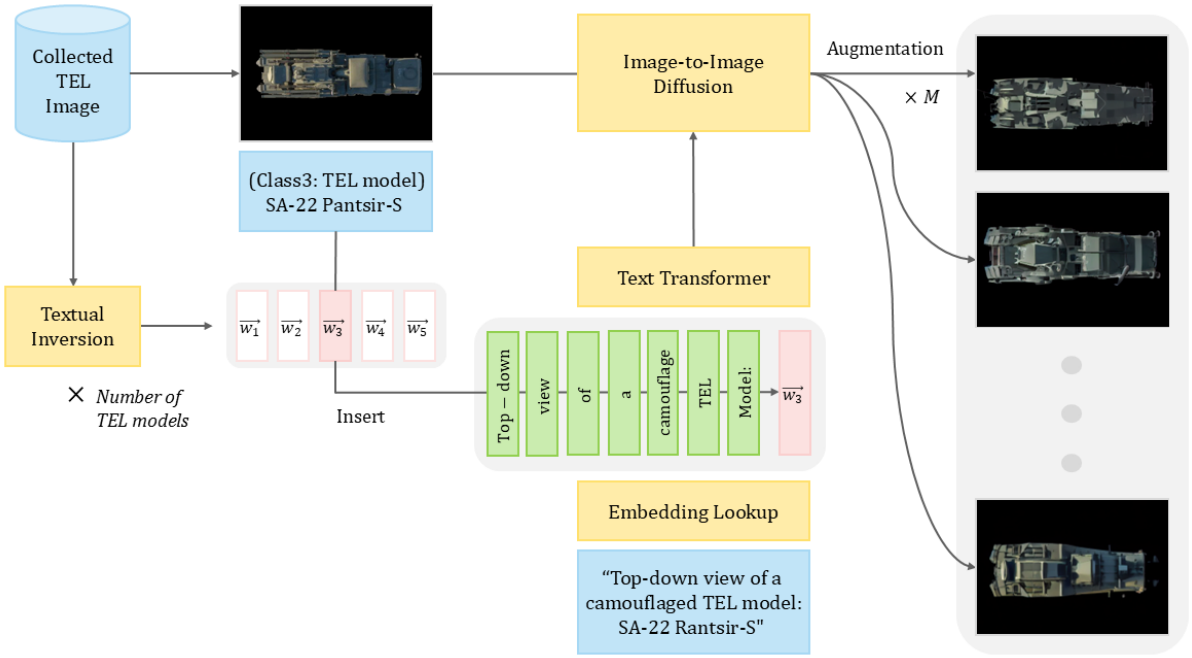

Figure 2. TEL Dataset Augmentation Process

The TEL dataset augmentation method proposed in this study is based on the DA-Fusion approach[13], which has demonstrated performance in previous research. This approach integrates two main components—Stable Diffusion and Textual Inversion—to create synthetic images suitable for the aerial and satellite imaging domains. Stable Diffusion, which relies on the Image-to-Image mechanism with text prompts, allows conditions such as "Top-down view," "camouflaged," and so on to be specified. Compared to alternative methods like GANs, it can generate a broader range of variations with a more natural appearance. This feature aligns well with our primary goals of maintaining the TEL's semantic features while achieving color, pattern, and angle diversity. Meanwhile, Textual Inversion trains text embeddings for TEL model names, enabling the Diffusion model to handle specialized military domain terminology (i.e., specific TEL models) that a general Diffusion model might not recognize. As a result, each TEL model's unique shape or design can be consistently maintained throughout the synthesis process. Therefore, DA-Fusion—combining Stable Diffusion and Textual Inversion—proves to be an effective method for generating large quantities of top-down

TEL images that capture domain-specific nuances, including those associated with military and satellite settings as well as distinctive camouflage patterns.

Figure 2 illustrates the overall framework of the proposed TEL dataset augmentation process, which comprises two primary stages: (1) Textual Inversion embedding training and (2) Image-to-Image Diffusion. In the data input stage, the collected TEL images undergo preliminary preprocessing, including background removal and basic cleaning. During the Textual Inversion embedding training stage, text embeddings corresponding to each TEL model name are learned, enabling the Stable Diffusion model to accurately reproduce or manipulate specific model types in response to appropriate text prompts. In the Image-to-Image Diffusion stage, both the original TEL image and a corresponding text prompt (e.g., "Top-down view of a camouflaged TEL Model {model class}") are provided as inputs. By adjusting parameters such as prompt tuning, noise levels, and iteration counts, a broad range of transformations is achievable, allowing significant variation while preserving the structural integrity of the TEL images. Specific details regarding parameter selection and evaluation methods are provided in the simulation section; here, the focus is on delineating the overall algorithmic flow.

3.2 TEL Object Masking and Insertion

This subsection proposes an object masking and insertion algorithm to seamlessly composite augmented TEL images into real satellite imagery. The previously generated TEL images often include a black background and TEL images can be captured (original TEL) or generated (augmented TEL) at varying angles and distances, resulting in different scales and rotations. Therefore, before compositing, it is crucial to extract only the object region and adjust its rotation and size. If the boundaries are not properly isolated, unnatural edges or artifacts may appear in the final composite image. Consequently, precise masking and contextually appropriate placement are essential for building a TEL detection dataset suited to military and security domains.

The detailed objectives for accurate TEL object masking are as follows. First, any unnecessary background within the TEL image must be removed with high precision. If boundary details are lost, the object's outline may break or overlap awkwardly with the military terrain during compositing, which directly degrades final image quality. Second, because TEL images can be collected at various scales and orientations, an automated procedure for angle correction and scale adjustment must be applied before compositing so that they match the resolution of the satellite image. Third, based on prior regional information for each satellite image, the TEL should be inserted at an appropriate location. By selecting strategic spots within the collected satellite images and compositing the TEL accordingly, contextual coherence can be maintained. Ultimately, the composite image and its corresponding object mask are stored together, allowing subsequent masking and inpainting procedures to ensure a natural and contextually relevant compositing process.

In this study, we employ an `OpenCV`-based approach that combines color-space thresholding and morphological operations to remove the black background from TEL images, followed by rotation correction and alpha blending to smoothly insert them into satellite images[21]. First, the black background can be easily isolated using color-space thresholds, after which a closing operation removes contiguous noise or small holes. Next, the TEL outline is detected through contour analysis, and the center coordinates and rotation angle are obtained using `minAreaRect` to align the TEL image. If necessary, a `margin_factor` parameter is introduced to handle margin space around the object. Finally, the TEL image is resized to match the resolution of the satellite image, and its alpha channel is adjusted to minimize boundary color bleeding.

The resulting TEL object is then inserted into the satellite image at coordinates and sizes determined by the relevant military context and satellite-image resolution. The composite image and its associated object mask are saved as separate files for subsequent inpainting or training of detection models. As shown in Algorithm 1, the proposed procedure can be broadly divided into four main phases: background removal (Step 1, Step 2), rotation and size adjustment (Steps 3, Step 4), boundary refinement(Step5), and final compositing (Step6). Moreover, the insertion-location selection logic, for example near tunnels or mountainous terrain, will be presented alongside the simulation results in Section 4, illustrating how this algorithm operates in real-world applications.

Algorithm 1 TEL Image Masking & Insertion

Step 1: Load Background & TEL
Load the background image and the TEL image.
Step 2: Remove Black Background
Identify and mask the background of the TEL image.
Apply a closing morphological operation.
Step 3: Contour & Rotation
Find the largest contour to identify the TEL region.
Obtain TEL angle and size information.
Rotate the TEL image to align it properly.
Step 4: Crop and Resize
Generate a bounding box and minimize margins to isolate the TEL object.
Resize the TEL image to match the target scale.
Step 5: Boundary Refinement
Remove any remaining background color bleeding.
Use erosion and thresholding.
Step 6: Insert into Satellite Image & Save
Paste the TEL image at the chosen coordinates.
Save the final composite image and mask.

3.3 Restoration of Masked Areas in Satellite Images

The direct integration of a TEL object into a satellite image can result in boundary discontinuities, leading to the introduction of synthetic artifacts. This occurs because the TEL object does not align seamlessly with the original satellite background, resulting in perceptible texture inconsistencies that affect the structural coherence of the composite image. Such artificial synthesis introduces discrepancies from real-world satellite data, potentially impacting the generalization performance of object detection

models and increasing susceptibility to false positives and false negatives. To mitigate this issue, it is essential to reconstruct the masked region to ensure cohesive integration between the inserted TEL object and the surrounding background, thereby maintaining texture continuity.

Therefore, this study proposes a method to reconstruct the surrounding region of the TEL object through inpainting by incorporating contextual features such as roads, terrain, and shadows in satellite images. To achieve this, the model optimizes texture coherence at object boundaries and ensures smooth spatial integration to enhance the structural consistency of the composite image. Additionally, the restoration process is designed to account for regional attributes and spatial dependencies present in real satellite imagery, facilitating accurate and consistent reconstruction.

To enhance texture consistency and structural coherence between inserted objects and their surrounding regions, this study adopts a Transformer-based inpainting framework. The method leverages a self-attention-driven architecture, which, in contrast to conventional CNN-based approaches, is capable of jointly modeling long-range dependencies and local spatial features. This modeling capacity is particularly suited for reconstructing satellite imagery that encompasses complex semantic and structural contexts. The restoration of the masked region is guided by the self-attention mechanism, which accounts for both localized pixel interactions and global feature patterns during reconstruction. This mechanism contributes to the mitigation of discontinuities at the interface between the inserted TEL object and the original satellite background, supporting a more consistent spatial integration. Moreover, the Transformer-based model demonstrates an ability to process high-resolution satellite imagery containing heterogeneous elements—such as built environments, natural terrain, and transportation infrastructure—thereby enabling boundary-level inpainting under diverse spatial conditions.

As illustrated in Figure 3, our study implements a Transformer-based restoration process that takes as input both a satellite image with an inserted TEL object and the corresponding masked regions surrounding the object. This approach segments the input masked region into pixel patches and employs the Self-Attention mechanism to simultaneously learn global and local information. This enables effective incorporation of both global context, such as roads and terrain where the TEL object is situated, and local context that defines the object's boundaries. The Transformer architecture is particularly advantageous for processing satellite imagery where contextual information is distributed across wide spatial ranges, as it effectively captures long-range dependencies. Our model employs multiple Transformer blocks to progressively fill the masked regions while reconstructing textures that maintain both spatial proximity and semantic consistency. To enhance restoration quality, we apply a Mask Updating technique that iteratively refines the restored regions, minimizing boundary discontinuities between TEL objects and their backgrounds. Furthermore, the integration of Mapping modules and Style Manipulation Module (SMM) preserves the inherent style and illumination characteristics of the satellite imagery while ensuring consistent texture generation. The final restored image is stored as a harmoniously synthesized composition where the TEL object is seamlessly integrated with the original satellite background, contributing to the development of a high-quality satellite image dataset for enhanced TEL detection capabilities. Detailed performance analysis of the restored images and various parameter adjustments will be thoroughly discussed in section 4 (Simulation).

The adopted Transformer-based inpainting framework is designed to preserve the semantic consistency of the TEL object while maintaining the structural fidelity of the original satellite imagery. To achieve this, we introduce an asymmetric weight modulation mechanism into the Transformer-based inpainting architecture, enabling the controlled adjustment of the relative contributions of the original and restored regions in the final composite output. This adjustment mechanism is formulated as the following components.

Original equation

$$I = I \cdot (1 - M) + I_{in} \cdot M \tag{1}$$

Modified equation

$$I = I \cdot (1 - M) \cdot \beta + I_{in} \cdot M \cdot (1 - \beta) \tag{2}$$

where, I is the final image, I_{in} represents the original satellite image, M is the binary masked region, and $\beta \in [0, 1]$ is a coefficient that regulates the relative influence of the restored and original regions. The proposed asymmetric weighting approach can be seamlessly integrated into the existing Transformer-based inpainting pipeline as a post-processing operation. By explicitly controlling the contribution of each source image, this method mitigates two common issues in generative restoration: over-reconstruction within masked areas and loss of semantic or structural consistency from the original image. Empirical experiments determined that the optimal range for β is between 0.2 and 0.4. For instance, when $\beta = 0.3$, the contribution of the Transformer-based restoration decreases to 30%, while the proportion of the original satellite image increases to 70%, preserving more of the original information. In general, maintaining 60–80% of the satellite background information yields improved structural coherence and perceptual consistency. This asymmetric weighting adjustment enables a balance between preserving scene-specific features and refining contextual transitions. Unlike conventional inpainting techniques that prioritize full restoration, the proposed approach emphasizes constrained integration, which better maintains semantic and geometric integrity. The effectiveness of the proposed weighting adjustment method is further analyzed in section 4 (Simulation) through simulation results.

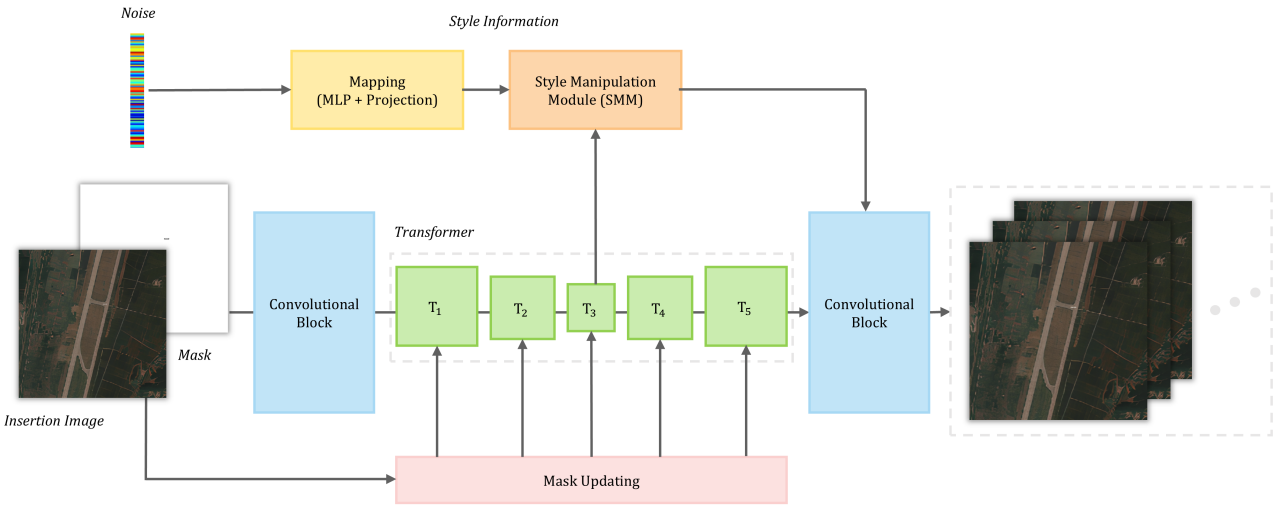

Figure 3. Inpainting Process of Missing areas in Satellite images using Transformer-based model

4 Simulation

4.1 Simulation of TEL Dataset Augmentation

This section provides a detailed explanation of how we implemented the DA-Fusion-based TEL image augmentation method in an actual simulation environment. First, we collected 58 images from 13 TEL models, then removed watermarks and backgrounds using GrabCut[22]. Next, we applied the DA-Fusion technique to the Stable Diffusion model (stable-diffusion-v1-4), learning new tokens for each model name via Textual Inversion (learning rate 5×10^{-4}, $1,000$ steps, Adam optimizer) and setting the input resolution to 512×512, 1,000 denoising steps, and a guidance scale of 7.5.

Table 1. Prompt Variants and Augmentation Outcomes

Version	SD-Guided Prompt	Key Outcomes
Default	`a photo of a {model}`	Model name appears as overlaid text; many images use side views.
Prompt V1	`Top-down view of a Transporter Erector Launcher (TEL), model name: {model}`	Side-view outputs drop sharply; colors become more varied and exteriors appear more ornate;
Prompt V2	`Top-down view of a camouflaged TEL (model: {model})`	Accurate top-down perspectives with realistic camouflage; structural details are preserved.

Since Stable Diffusion is sensitive to textual prompts, we experimented with various prompts and summarized the results in Table 1. Ultimately, we adopted the prompt "Top-down view of a Transporter Erector Launcher, model name: {model}" with keywords like "top-down view" and "camouflaged" to enhance military TEL image suitability. Figure 4 presents the images generated by each prompt.

Figure 5 shows how semantic invariance is maintained across different TEL models. These augmented results preserve essential semantic features—such as vehicle shape and missile mounting structure—while allowing variations in camouflage, color, and lighting. Thus, the top-down perspective remains intact, confirming that our method achieves 'semantic invariance' in a qualitative sense.

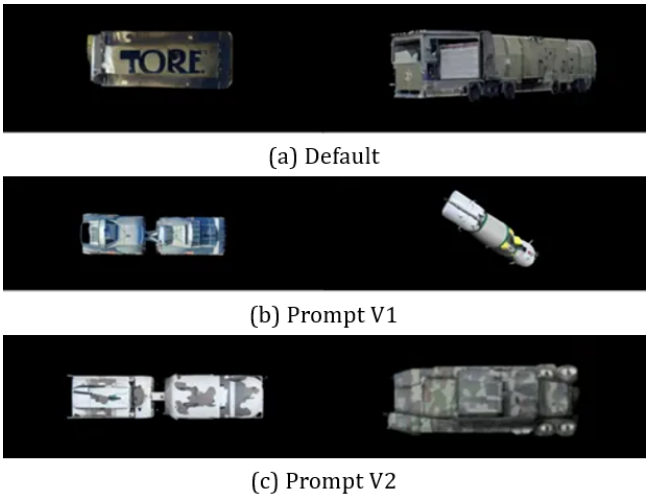

Figure 4. Augmentation Results by Prompt

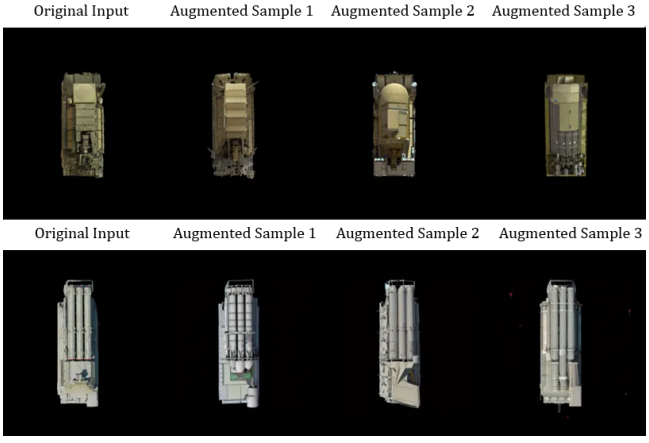

Figure 5. Diverse Augmentations with Semantic Invariance

To compare methods under identical input conditions, we evaluated our diffusion-based DA-Fusion against a WGAN-based approach. Quantitatively, we used the Fréchet

Inception Distance (FID), treating real and generated images as high-dimensional Gaussians to measure distributional distance. We formed 100 pairs of original images by randomly sampling the 58 TEL images at a one-to-one ratio and calculated the threshold as the 95th percentile of the resulting FID distribution, 158.62. The DA-Fusion method achieved an average FID of 148.16, below this threshold, while the WGAN approach recorded 243.52, exceeding it. Figure 7 shows diffusion-based outputs looking more natural and richly detailed than GAN-based ones. These quantitative and qualitative results together demonstrate that diffusion-based augmentation significantly outperforms the GAN-based method.

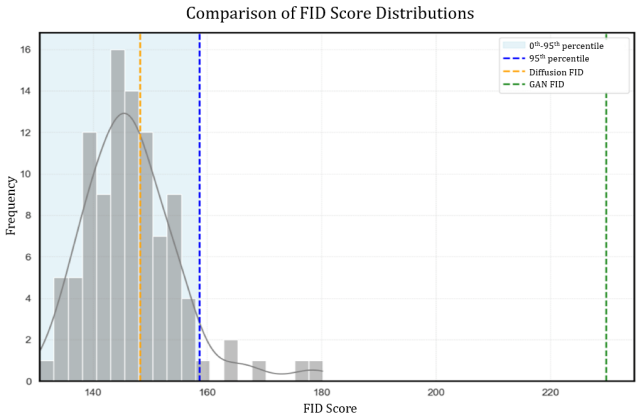

Figure 6. Quantitative Evaluation: Diffusion vs GAN

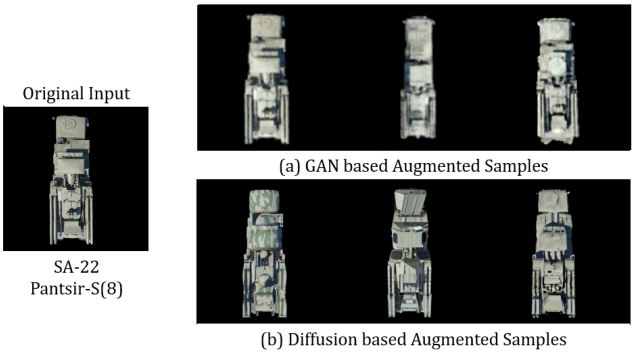

Figure 7. Qualitative Evaluation: Diffusion vs GAN

4.2 *Simulation of TEL Object Masking and Insertion*

In the previous Section 4.1, we obtained both the original TEL and the augmented TEL images; to seamlessly composite these into real satellite imagery, a precise object-masking step is essential. In this simulation, following Section 3.2, we used the OpenCV library[21] to remove background via color thresholding (`cv2.inRange`), extract TEL contours with `findContours`, and normalize object orientation using `minAreaRect`. We then applied morphological erosion, `erode`, to suppress boundary noise and used `warpAffine` to rotate and align each object. Figure 8 illustrates examples of the resulting TEL objects alongside their binary masks, confirming that the background is effectively eliminated and the TEL shapes are accurately isolated.

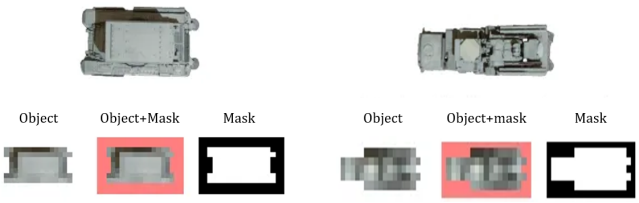

Figure 8. TEL Object Masking Simulation Results

Once the masking is complete, the next key step is to select insertion sites that reflect plausible military scenarios. We collected satellite imagery from strategic locations—Sunan, Geumsu Mountain, Haeju, and Rason—and identified candidate areas with high potential for TEL concealment and launch. In particular, Sunan Airport in Pyongyang and the tunnel entrances around Geumsu Mountain have documented missile-operation histories, making them ideal for realistic data synthesis. Because TEL deployments require both covertness and launch readiness, terrain features such as mountain slopes, tunnel portals, and nearby defensive structures were chosen as primary insertion candidates. Figure 9 shows examples of TEL placement near Sunan Airport—where roughly 18 missile launches took place during 2023—and at nearby tunnels Mount Kumsu, illustrating our construction of realistic synthetic data..

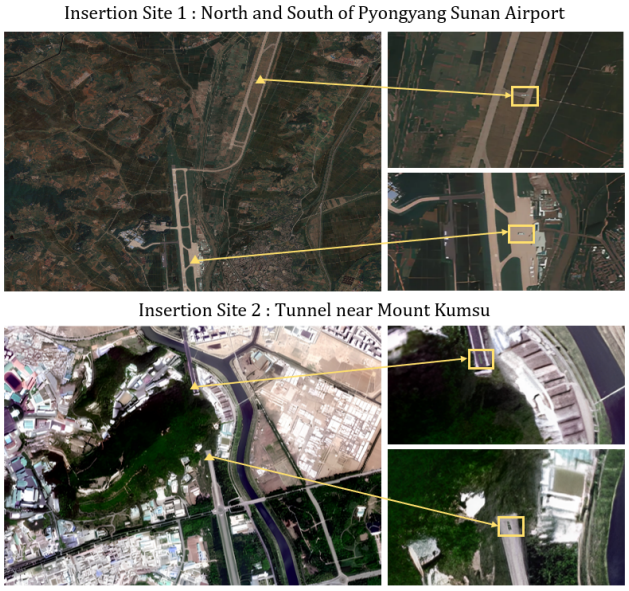

Figure 9. Examples of TEL insertions overlaid on real satellite imagery at Sunan Airport and a tunnel near Mount Kumsu, both within the Pyongyang area.

4.3 *Simulation of Masked Area Restoration in Satellite Images*

In this subsection, an inpainting procedure is applied to the masked region surrounding the TEL object, based on the insertion results described in Section 4.1. The inserted image reflects both the TEL object and its surrounding masked area, incorporating the actual geographic context of the satellite imagery. A modified version of the Transformer-based model architecture is employed for the restoration process. The model requires, as input, a satellite image

(a) Insertion site (b) Inserted object (c) Restoration Result

Figure 10. Inpainting Results using Transformer Models

with the inserted TEL object and its corresponding binary mask. The experiments utilize the pre-trained CelebA-HQ-256 model, which is trained on a large-scale high-resolution image dataset. This model is selected due to its capacity to handle discontinuities in texture boundaries, leveraging learned features that are applicable to satellite imagery with complex spatial patterns. The restoration results indicate that the masked regions adjacent to the TEL object were reconstructed in a manner that aligns with the surrounding background textures. This behavior is attributed to the self-attention mechanism in the Transformer architecture, which concurrently considers local pixel relationships and global context. Figure 10 illustrates a representative case in which the interior and exterior regions of the TEL object are independently processed, and the boundary mask is restored to achieve texture continuity with neighboring regions. Even in instances where the TEL insertion produces noticeable discontinuities at object boundaries, the model is able to recover a visually continuous transition in the reconstructed output.

The restoration results obtained using the modified Transformer-based model demonstrate that, in most cases, the model implicitly separates the interior and exterior regions of the TEL object, resulting in independent reconstruction behavior along object boundaries. However, in certain instances, the inpainting process extends internal visual characteristics of the TEL object into the surrounding background. This phenomenon is particularly observed when the TEL object contains complex textures or visually salient attributes, such as high chromaticity or strong edge contrast. To address this issue, an asymmetric weight adjustment technique was employed to regulate the relative contributions of the original image and the Transformer-based restoration. This technique increases the influence of the original satellite image in the final output, thereby reducing the unintended propagation of internal object features into adjacent regions. Figure 11 presents a visual example of this adjustment, demonstrating that increasing the contribution of the original image can result in more stable and spatially coherent restoration outcomes, particularly at object boundaries where undesired feature diffusion occurs.

The example shown in Figure 12 of the final TEL detection dataset includes each TEL model name and insertion case at various locations, together with a range of adjustment parameters such as x-fold scaling. In Figure 12, Model A and Model B were inserted into the northern and southern areas of Sunan Airport, respectively, and the insertion coordinates, rotation angles, and scale factors are all annotated to simulate a realistic satellite-image environment. This simulation systematically incorporates domain knowledge to account for the various scenarios in which TEL might be observed in satellite imagery such as scaling, rotation, positioning, etc. Note that the images presented here are the raw outputs before any post-processing; depending on the type of satellite sensor and the observed wavelength band, additional color adjustment or filtering post-processing steps may be applied in this study, we first augmented 58 TEL images via diffusion and then constructed the TEL detection dataset over N regions. The average time required to augment a single image at each stage is as follows (see Table 2):

Stage	Average Time
Diffusion-based TEL augmentation	2.76 seconds
Insertion and adjustment	0.319 seconds
Transformer-based inpainting	2.15 seconds

Table 2. Average per-image processing time for each pipeline stage on augmented data for TEL detection

Building the TEL detection dataset of 580 images required a total of 50 minutes 32.82 seconds, corresponding to an average of 5.229 seconds per augmentation.

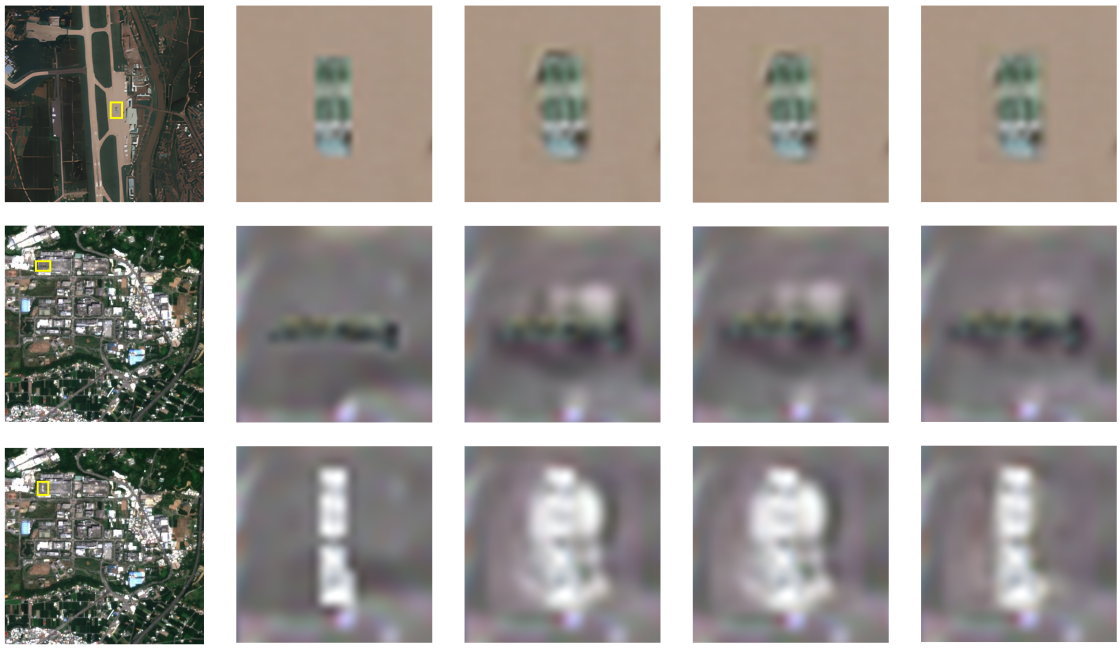

Figure 11. Comparison of Experimental Results on Asymmetric Weight Adjustment

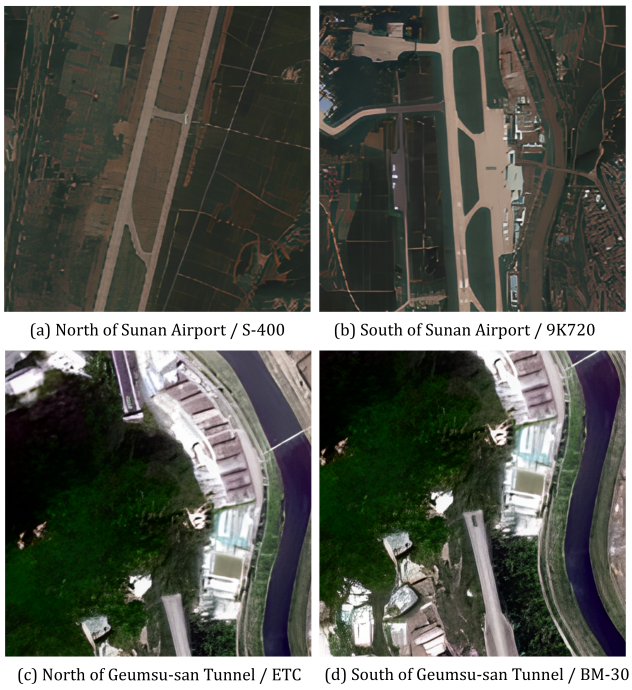

Figure 12. Examples of the final TEL dataset generated by the full pipeline near Sunan Airport and Mount Kumsu in Pyongyang

5 Conclusion

This paper introduces a three-stage synthetic-data pipeline that tackles severe data scarcity in TEL detection from satellite imagery. The pipeline integrates semantic-invariant diffusion augmentation, context-aware insertion, and transformer-based boundary inpainting into a unified workflow. It produces high-fidelity composites that preserve both object geometry and terrain constraints. Against a WGAN baseline, our dataset cuts the Fréchet Inception Distance by 38 %. Qualitative evaluations demonstrate diverse camouflage patterns, realistic military-site placements, and seamless boundary repairs. TEL detection remains under-explored but strategically vital. Our approach significantly advances detector training and is expected to reinforce broader security and early-warning capabilities. Two key challenges remain. We still rely on manual selection of insertion coordinates, and the pipeline currently handles only RGB imagery. To address these issues, we will automate site selection using large-scale, military-context geospatial retrieval. We also plan to extend the pipeline to non-optical imagery and to synthesise background satellite data that reflect realistic spatio-temporal environmental variations.

References

1. Zhai Y, Ma H, Cao H et al. Mf-sarnet: Effective cnn with data augmentation for sar automatic target recognition. *The Journal of Engineering* 2019; 2019(19): 5813–5818. DOI:https://doi.org/10.1049/joe.2019.0218. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/joe.2019.0218. https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/joe.2019.0218.
2. Marcum RA, Davis CH, Scott GJ et al. Rapid broad area search and detection of Chinese surface-to-air missile sites using deep convolutional neural networks. *Journal of Applied Remote Sensing* 2017; 11(4): 042614. DOI:10.1117/1.JRS.11.042614. URL https://doi.org/10.1117/1.JRS.11.042614.
3. Klein MD, Edel ZJ, Packard CD et al. Automatic detection and identification of ground vehicles with YOLO-based deep learning using MuSES-generated OPIR synthetic imagery. In Bouma H, Prabhu R, Yitzhaky Y et al. (eds.) *Artificial Intelligence for Security and Defence Applications II*, volume 13206. International Society for Optics and Photonics, SPIE, p. 132060N. DOI:10.1117/12.3031663. URL https://doi.org/10.1117/12.3031663.
4. Zhang L, Zheng J, Li C et al. Ccdn-detr: A detection transformer based on constrained contrast denoising for multi-class synthetic aperture radar object detection. *Sensors* 2024; 24(6). DOI:10.3390/s24061793. URL https://www.mdpi.com/1424-8220/24/6/1793.
5. Koopman P, Fickinger A and Sinha A. Synthetic data generation for ai training and evaluation: A review. *arXiv preprint arXiv:240407503* 2024; URL https://arxiv.org/abs/2404.07503.
6. Khammari S, Sánchez-Biezma EF, Sukhanov S et al. Synthetic data augmentation for earth observation object detection tasks. In *Workshop on Machine Learning for Remote Sensing (ML4RS) at ICLR 2024*. pp. 1–1. URL https://iclr.cc/virtual/2024/22020. Poster.
7. II ABC, Davis CH, Scott GJ et al. Broad area search and detection of surface-to-air missile sites using spatial fusion of component object detections from deep neural networks, 2020. URL https://arxiv.org/abs/2003.10566. 2003.10566.
8. Arjovsky M, Chintala S and Bottou L. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, pp. 214–223.
9. Karras T, Laine S and Aila T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4401–4410.
10. Ho J, Jain A and Abbeel P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., pp. 6840–6851. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.
11. Rombach R, Blattmann A, Lorenz D et al. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10684–10695.
12. Gal R, Alaluf Y, Atzmon Y et al. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:220801618* 2022; .
13. Trabucco B, Doherty K, Gurinas M et al. Effective data augmentation with diffusion models, 2023. URL https://arxiv.org/abs/2302.07944. 2302.07944.
14. Ghiasi G, Cui Y, Srinivas A et al. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 2918–2928.
15. Zhu P, Abdal R, Femiani J et al. Barbershop: Gan-based image compositing using segmentation masks. *arXiv preprint arXiv:210601505* 2021; .
16. Quan W, Chen J, Liu Y et al. Deep learning-based image and video inpainting: A survey. *International Journal of Computer Vision* 2024; 132(7): 2367–2400.
17. Criminisi A, Perez P and Toyama K. Object removal by exemplar-based inpainting. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2. IEEE, pp. II–II.
18. Yu J, Lin Z, Yang J et al. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 5505–5514.
19. Nazeri K, Ng E, Joseph T et al. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*. pp. 0–0.
20. Li W, Lin Z, Zhou K et al. Mat: Mask-aware transformer for large hole image inpainting, 2022. URL https://arxiv.org/abs/2203.15270. 2203.15270.
21. Bradski G. The opencv library. *Dr Dobb's Journal: Software Tools for the Professional Programmer* 2000; 25(11): 120–123.
22. Rother C, Kolmogorov V and Blake A. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)* 2004; 23(3): 309–314.