Top-Down Data Generation Using Vision–Language Models for Object Detection in Constrained Battlefield Environments

Yerin Kim*†, Jaeuk Baek*, Donggyu Choi*, and Chang-eun Lee*†
*Digital Convergence Research Laboratory, ETRI, Daejeon 34129, Korea
†University of Science and Technology, Daejeon 34113, Korea
{imyeslin, jubaek, dgchoi, celee}@etri.re.kr

***Abstract*—In this paper, we propose a novel object detection framework based on Vision–Language Models (VLMs). The framework integrates synthetic top-down image generation, automatic object labeling, and efficient composition of real and synthetic data for robust model training. Specifically, ground-view data are collected using mobile robots, and VLMs are employed to transform this data into high-quality synthetic top-down images. We optimize both positive and negative prompts in VLMs to maximize image quality. The generated outputs are then combined with real top-down images to train the object detection model. To automate object labeling, we utilize the Grounding DINO model, a vision–language object detector, and apply data augmentation techniques to improve generalization. Through a series of experiments, we investigate the impact of different synthetic-to-real image ratios in the training set and identify the optimal combination through quantitative performance analysis. Experimental results demonstrate the effectiveness of the proposed framework in generating high-quality synthetic top-down view data and in enhancing object detection performance in constrained battlefield environments.**

***Index Terms*—Vision-Language Models (VLMs), top-down images generation, object detection, synthetic data, battlefield environments**

I. INTRODUCTION

Compared to ground-view images, top-down images offer a wider field of view and an unobstructed line-of-sight (LOS) perspective, which can significantly improve both situational awareness and object detection. However, in indoor battlefield environments where CCTV infrastructure is absent and aerial platforms cannot be deployed, acquiring top-down images remains extremely challenging. Recent advances in Vision–Language Models (VLMs) [1] enable the synthesis of realistic top-down images from readily available ground-view data, thereby reducing the dependence on large-scale manual data collection. In this paper, we propose a VLM-based approach that generates synthetic top-down images from ground-view inputs and integrates them with real images in varying proportions for model training. We systematically evaluate object detection performance to rigorously assess the practical feasibility and effectiveness of incorporating synthetic top-down images into training pipelines under constrained battlefield conditions.

II. RELATED WORK

In this section, we review a series of techniques for generating synthetic data, focusing on their applicability and performance. Particular emphasis is placed on their feasibility and robustness in constrained environments.

A. Image to 3D

- **Stable-Point-Aware-3D** [2]: This technique improves geometric accuracy by aligning multi-view features, but the generated point clouds occasionally exhibit artifacts such as scattered points. Furthermore, extracting precise color and material information from a single image remains challenging. When applied to full-length images, environmental factors such as smoke, dust, and lighting further exacerbate these instabilities, consequently reducing the reliability of the reconstruction results.
- **PartPaker** [3]: Objects are segmented into part-level units to facilitate reconstruction, with segmentation accuracy critically depends on input data quality. Hidden areas and indistinct boundaries pose substantial challenges to accurate segmentation. In complex environments such as battlefields, partial occlusions and damaged structures often result in incomplete or distorted object forms, increasing the risk of segmentation errors.
- **Instant Mesh** [4]: High-quality meshes can be rapidly generated from a single image. Performance declines significantly in complex or occluded scenes. Achieving more accurate results requires images obtained from multiple viewpoints. In battlefield environments, where entities such as ground robots are present, relying only on single-view input makes reliable reconstruction of the complete scene structure challenging.
- **FLUX.1-Kontext-dev** [5]: By leveraging a diffusion-based virtual camera framework, this method performs prompt-based synthetic view generation, producing high-quality perspective transformations such as top-down views from single ground-level images. It effectively maintains visual consistency in terms of object scale, orientation, and illumination, thereby making it highly suitable for synthetic data generation and augmentation in constrained environments.

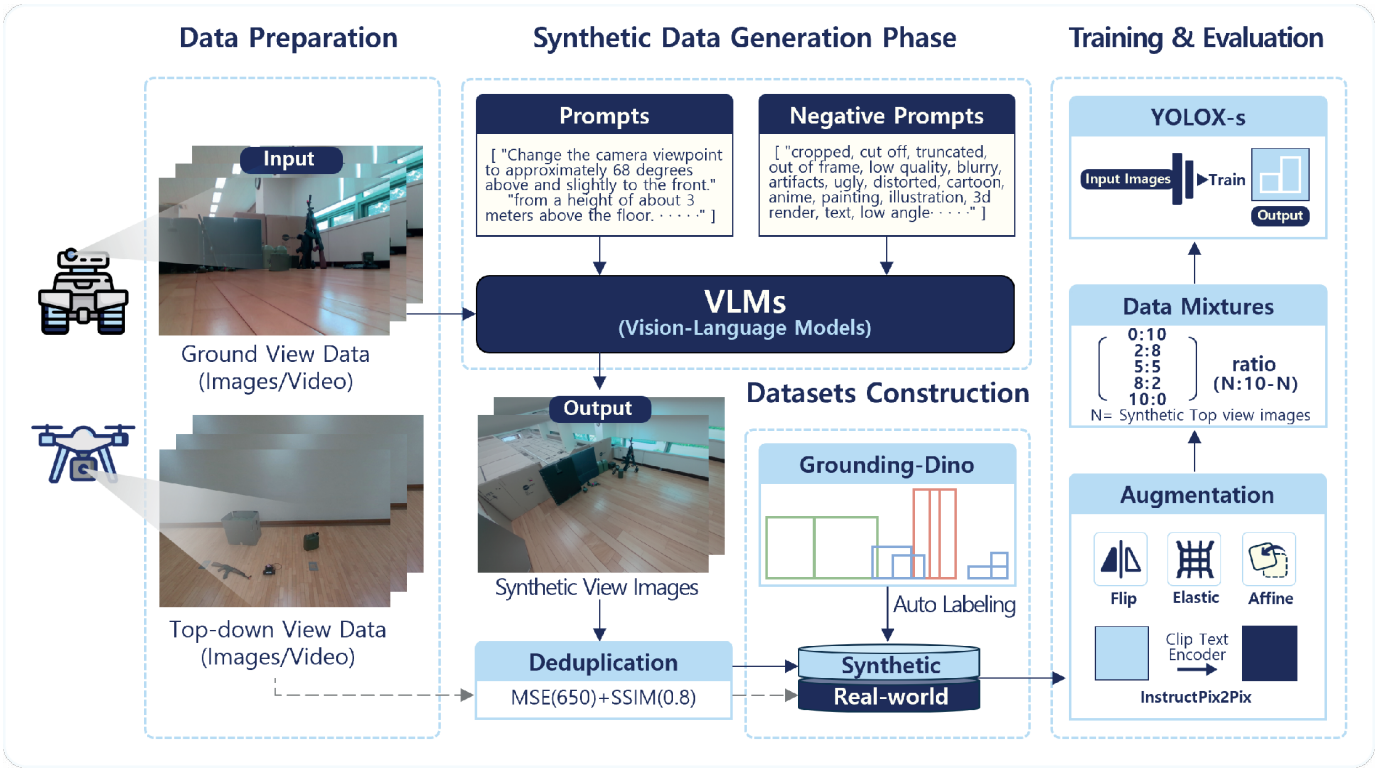


Fig. 1. Overall Process of the Proposed VLMs-Based Synthetic Data Generation and Object Detection Pipeline.

In summary, existing image-to-3D techniques have strengths and limitations. Stable-Point-Aware-3D [2] enhances geometric alignment but is vulnerable to noise, while PartPaker [3] enables part-level modeling but suffers under occlusion. Instant Mesh [4] can rapidly generate meshes from single images, though it generally requires multi-view inputs. FLUX.1-Kontext-dev [5] performs prompt-driven synthetic view generation, producing top-down views from ground-level images. These characteristics highlight the need for more robust approaches when applying synthetic data generation in constrained battlefield environments.

B. Stable Virtual Camera

A viewpoint transformation framework operates directly on 2D images with depth or feature guidance [6]. It preserves object scale, spatial arrangement, and texture fidelity. Multiple images are required for generating top-down views without full 3D reconstruction. When the disparity between the input and target viewpoint is large, results may degrade. In scenes with many complex objects, viewpoint transformation can cause spatial and boundary errors.

C. Vision-Language Model (VLMs)

Vision-Language Models [1] leverage multimodal inputs integrating visual and textual information to perform generative tasks. This paradigm enables the transformation of ground-view images into synthetic top-down views while preserving scene structure. As it does not rely on multi-view inputs, this approach is suitable for constrained battlefield environments.

III. PROPOSED FRAMEWORK

The proposed framework consists of three stages: synthetic data generation, dataset construction, and model training and evaluation, as summarized in Figure 1.

A. Synthetic Data Generation

We adopt the VLM, i.e., FLUX.1-Kontext-dev [5], to generate synthetic top-down view images, where various prompts and negative prompts are analyzed and used to convert ground-view images into top-down views. To do this, we assume the virtual camera configured at a height of 3 m above the ground with an elevation angle of approximately 68°, while zooming or reframing is prohibited. A unified prompt template was applied to enforce global constraints on camera viewpoint and scene fidelity:

> Change the camera viewpoint to approximately 68° elevation, slightly tilted forward, at a height of approximately 3 m above the floor. The environment consists of a room with a ceiling height of about 3.6 m and a standard door height of approximately 2.0 m; floor tiles measure roughly 60 cm each. A 24 mm-equivalent wide-angle field of view without zoom was used. The entire floor area is captured within the frame without cropping or reframing. All existing objects were preserved in their original positions, sizes, orientations, illumination conditions, and textures. No new objects were introduced during image synthesis.

TABLE I
PROMPTS AND NEGATIVE PROMPTS FOR PRESERVING ORIGINAL OBJECT PROPERTIES IN SYNTHESIZED TOP-DOWN VIEW IMAGE GENERATION.

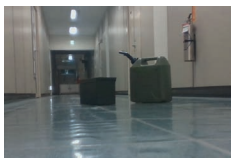
Object	Input	Output	Prompts	Negative Prompts
Object 1			"Object 1 in a top-down perspective, keeping its size, position, and layout as in the original scene."	"No added grenades, no added mines, no duplication, no truncation, no hallucinated Object 1."
Object 2			"Object 2 in a top-down perspective, keeping its shape, orientation, and position as in the original scene."	"No added guns, no added Object 2, no scopes, no bipods, no duplication, no truncation."
Object 3			"Object 3 (e.g., jerry cans, chemical drums) in a top-down perspective, keeping their number, scale, and layout as in the original scene."	"No added containers, no added drums, no chemical symbols, no duplication, no truncation."
Object 4			"Object 4 (e.g., sandbags, barriers) in a top-down perspective, keeping their arrangement and appearance as in the original scene."	"No added sandbags, no added barriers, no duplication, no rearrangement, no truncation."
Object 5			"Object 5 in a top-down perspective, keeping its size, position, and alignment with the wall as in the original scene."	"No added Object 5, no added windows, no hallucinated handles or locks, no truncation, no duplication."
Object 6			"Object 6 clearly on the floor from an oblique top-down perspective, keeping its appearance and layout as in the original scene."	"No Object 6 with three wheels, no Object 6 with four wheels, no multi-wheeled Object 6, no many-wheeled Object 6."

Table I summarizes the object-specific prompt variations for six target objects, and each prompt is designed to preserve its original properties. The generated images were subsequently integrated with real-world data in varying ratios for performance evaluation. To ensure reliability, Mean Square Error (MSE) (≤ 650) [7] and Structural Similarity Index Measure (SSIM) (≥ 0.8) [8] thresholds were applied to filter out duplicate and low-quality samples.

B. Dataset Construction and Augmentation

In this section, a training dataset is created by integrating synthetic and real images. Specifically, synthetic images generated from VLM model are automatically labeled through automatic annotation with Grounding DINO [9]. Approximately 1,800 images were augmented to improve robustness using Albumentations [10] (horizontal flip, elastic and affine transforms) and InstructPix2Pix [11] (night and dust effects).

C. Model Training with Synthetic–Real Ratios

In the final phase of our framework, the object detection model is trained using various synthetic-to-real image ratios (0:10, 2:8, 5:5, 8:2, 10:0). For each ratio, a total of 10,000 samples are utilized. We assess the extent to which the generated synthetic top-down images can effectively complement real data in the training process.

IV. EXPERIMENTS AND RESULT

A. Experimental Setup

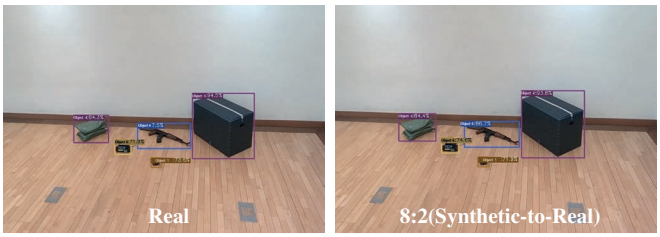
All experiments were conducted under a consistent hardware and software environment with an NVIDIA RTX 5070 Ti GPU, Ubuntu 20.04 LTS, CUDA 12.8, and Python 3.9. The object detection framework was implemented using YOLOX-S [12] as the baseline detector and trained for 300 epochs. Training was performed with a batch size of 16.

TABLE II
OBJECT DETECTION PERFORMANCE (MAP) ACROSS SYNTHETIC-TO-REAL DATA RATIOS

Ratio (Synthetic:Real)	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0:10	82.5	99.0	94.0	70.7	75.3	87.7
2:8	72.9	90.1	82.4	38.8	62.5	75.6
5:5	70.4	89.6	83.0	67.8	68.0	80.8
8:2	72.5	90.1	82.0	68.4	67.9	75.4
10:0	88.5	98.6	96.2	65.0	80.0	88.4

TABLE III
QUALITATIVE RESULTS OF MODELS TRAINED WITH VARIOUS SYNTHETIC-TO-REAL DATA RATIOS

Similar Environment



Dissimilar Environment

B. Quantitative Evaluation

Quantitative results are summarized in Table II, showing object detection performance across models trained with varying synthetic-to-real (S:R) ratios. Similar performance is observed for most metrics except small object mAP. Using only real data (0:10), the model achieved an mAP of 82.5 with AP50 and AP75 of 99.0 and 94.0, respectively, while small object accuracy was lower ($AP_S = 70.7$). Using only synthetic data (10:0) resulted in the highest overall mAP (88.5) and large-object accuracy ($AP_L = 88.4$), but weaker small-object performance ($AP_S = 65.0$). Balanced ratios (5:5, 8:2) yielded stable results across scales, improving small object accuracy (AP_S up to 68.4).

This result suggests that while synthetic data can effectively complement real data, the proportion of small objects in the synthetic dataset was relatively low. This imbalance likely reduced the model's exposure to small-object patterns during training, leading to degraded small-object detection performance. These findings highlight the importance of controlling object-size distributions when generating synthetic data for object detection tasks.

C. Qualitative Evaluation

We compare the object detection results of models trained with real data only (left) and with a mix of synthetic and real data (8:2 ratio, right) in both similar and dissimilar test environment. Both models exhibit comparable performance in similar environment, whereas the model trained with the mixed dataset demonstrates superior detection performance in dissimilar environment. As shown in Table III, the inclusion of synthetic data exposes the model to a broader range of variations in lighting conditions, background clutter, and object orientations. This increased exposure enables the model to maintain robust detection performance even under challenging battlefield-like conditions, such as the cluttered backgrounds and varying object positions illustrated in the dissimilar environment examples.

The diversity introduced by synthetic data helps the model generalize better to unseen scenarios by simulating conditions that are not fully represented in the real dataset. Incorporating synthetic images enhances the model's robustness and reduces overfitting to specific environmental factors.

V. CONCLUSION

We proposed a Vision–Language Model (VLM)-based preprocessing framework [1] for generating synthetic top-down view data for unmanned systems operating in constrained battlefield environments. Using optimized prompts and negative prompts [5], we generated synthetic top-down images and combined them with real images captured by a tripod to train YOLOX-based object detection models [12]. Experiments showed that mixing synthetic and real data improved detection performance across dissimilar environments. The method also maintained stable detection performance under data-scarce conditions, demonstrating the practical value of synthetic data in augmenting limited real-world datasets. This study highlights the potential of prompt-driven synthetic data generation as an effective means of reducing the reliance on costly and time-consuming real data collection. The results demonstrate that synthetic data can complement real data without degrading performance, even when real data are limited, thereby supporting the scalability of object detection systems for battlefield-related applications. Furthermore, the proposed framework can serve as a flexible preprocessing pipeline that can be adapted to various types of visual data and extended to other perception tasks beyond object detection.

In future work, we will evaluate our framework under more diverse battlefield-like conditions, including outdoor occlusions, dynamic lighting, and cluttered or congested scenes. We will also introduce data augmentation techniques specifically tailored for small objects to mitigate the observed performance degradation in small-object detection. Additionally, we plan to explore multi-scale training strategies and incorporate fine-grained supervision to further enhance small-object detection performance. These efforts will help establish a more robust and scalable pipeline for reliable object detection in real-world battlefield environments.

ACKNOWLEDGMENT

This work was supported by Korea Research Institute for defense Technology planning and advancedment(KRIT) grant funded by the Korea government(DAPA(Defense Acquisition Program Administration)) (No. 21-107-E00-009-02, “Realtime complex battlefield situation awareness technology”)

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[2] Z. Huang, M. Boss, A. Vasishta, J. M. Rehg, and V. Jampani, “Spar3d: Stable point-aware reconstruction of 3d objects from single images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[3] J. Tang, R. Lu, Z. Li, Z. Hao, X. Li, F. Wei, S. Song, G. Zeng, M.-Y. Liu, and T.-Y. Lin, “Efficient Part-level 3D Object Generation via Dual Volume Packing,” *arXiv preprint arXiv:2506.09980*, 2025.

[4] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan, “Instantmesh: Efficient 3D Mesh Generation from a Single Image with Sparse-view Large Reconstruction Models,” *arXiv preprint arXiv:2404.07191*, 2024.

[5] B. F. Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne, T. Dockhorn, J. English, Z. English, P. Esser *et al.*, “Flux. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space,” *arXiv preprint arXiv:2506.15742*, 2025.

[6] J. J. Zhou, H. Gao, V. Voleti, A. Vasishta, C.-H. Yao, M. Boss, P. Torr, C. Rupprecht, and V. Jampani, “Stable Virtual Camera: Generative View Synthesis with Diffusion Models,” *arXiv preprint arXiv:2503.14489*, 2025.

[7] W.-H. Chen and W. Pratt, “Scene adaptive coder,” *IEEE Transactions on Communications*, vol. 32, no. 3, pp. 225–232, 1984.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[9] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, “Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection,” in *European conference on computer vision*. Springer, 2024, pp. 38–55.

[10] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, “Albumentations: Fast and Flexible Image Augmentations,” *Information*, vol. 11, no. 2, p. 125, 2020.

[11] T. Brooks, A. Holynski, and A. A. Efros, “InstructPix2Pix: Learning to Follow Image Editing Instructions,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18 392–18 402.

[12] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, “YOLOX: Exceeding YOLO Series in 2021,” *arXiv preprint arXiv:2107.08430*, 2021.