A Study on Military Object Detection in Panoramic View Using Stable Diffusion

Jiwon Lee
Telecommunications Media Research Laboratory (ETRI)
Daejoen, Korea
ez1005@etri.re.kr

Sungwon Moon
Telecommunications Media Research Laboratory (ETRI)
Daejoen, Korea
moonstarry@etri.re.kr

Dowon Nam
Telecommunications Media Research Laboratory (ETRI)
Daejoen, Korea
dwnam@etri.re.kr

***Abstract*— The bottleneck in the development of military object detectors based on deep learning is the costly and high-difficulty training data collection. Although various methods have been proposed to solve this problem, augmentation methods for sparse data, which is extremely difficult to collect, is still an area of high difficulty that requires continuous research. This paper proposes a method to generate training data of an object detector for military smart glasses with a panoramic view with extremely insufficient learning data by using outpainting with stable diffusion. Through experiments, it has been proved that the proposed method has far superior training performance than training data generated by simply cropping or concatenating existing training data in the form of a panoramic view. The proposed method can be used as one of the excellent data augmentation methods in a situation where an image of unusual size, such as a fish-eye view image or an ultra-wide-angle image, is required as training data.**

Keywords—military object detection, military smart glasses, panoramic view generation, stable diffusion outpainting

I. Introduction

Deep learning object detection technology has moved beyond the academic level to a level that can be applied in various real-world fields. In the military, opinions are being gathered to accept the superior technology of the industry in the area of surveillance, which was previously handled by manpower, and attempts are continuously being made to graft deep learning-based object detection technique.

Currently, the Korean military is preparing to deploy a soldier wearable surveillance tool in the form of AR glasses similar to Microsoft's IVAS (Integrated visual augmentation system) [1], which is to be deployed in the US Army. In order to increase the usability of the planned wearable smart glasses for soldiers, a function that automatically detects objects to be monitored based on an input image and provides information about the detected objects is essential. However, the image input from the smart glasses is in the form of a panoramic view in which the horizontal size is 2 to 3 times larger than the vertical size, and performance degradation inevitably occurs when an existing object detector that inputs a standard-sized image is directly applied. That is, to develop an object detector that works well in smart glasses, a deep learning model must be trained with training data in the form of a panoramic view similar to the input image.

In order to develop deep learning-based technology, a certain amount of training data is essential, and a lot of money and time must be invested. However, in military surveillance, training data is sufficient for common surveillance targets such as people and vehicles, but not for targets that are important but infrequent surveillance targets such as warships, tanks, and fighter jets. Moreover, in the case of panoramic images, there is almost no accumulated data, so technology development is impossible with the collected data.

There are two main ways to solve the above problem in a traditional way without additional data collection. The first method is to use the existing training data as training data after cropping the image to fit the panoramic view. However, when using this method, the size of the object to be detected becomes relatively large, so the deep learning model loses the opportunity to train a relatively small object. Another method is to connect two images horizontally to create a panoramic view for training. However, there is a problem in that the boundary between the two images is difficult to connect naturally, and a solution is needed when an object exists on the boundary.

As a way to solve the above problem, this paper proposes a method to generate training data for object detectors for smart glasses from existing military actual data using the outpainting with stable diffusion model(referred to as SD). This method is a way to obtain a panoramic view without damaging the existing training data by using a generation model. We train the object detector using the data augmented by this method, and try to prove the validity of the proposed method through performance verification.

The structure of this paper is as follows. Section 2 briefly introduces the SD to be used in this paper, and then shows the aspect of military training data in the form of a panoramic view generated by outpainting with SD in Section 3. The validity of the training data generated by the experimental results is verified in Section 4, and a conclusion is drawn in Section 5.

II. Outline of Stable Diffusion

SD was developed with support from Stability AI and Runway ML based on High-Resolution Image Synthesis with Latent Diffusion Models [2] of the Machine Vision & Learning Group (CompVis) laboratory at the University of Munich, Germany. It is a text-to-image artificial intelligence model distributed by Stability AI under an open source license.

The architecture of the SD model is shown in Figure 1. The model essentially consists of three artificial neural networks: CLIP, UNet, and VAE (Variational Auto Encoder). When a user inputs text, the text encoder (CLIP) converts the user's text into a language that UNet can understand, called a token, and UNet denoises randomly generated noise based on the token. Repeated denoising produces a proper image, and it is VAE's job to convert this image into pixels.

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

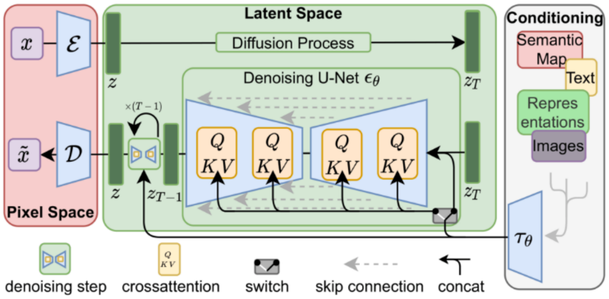
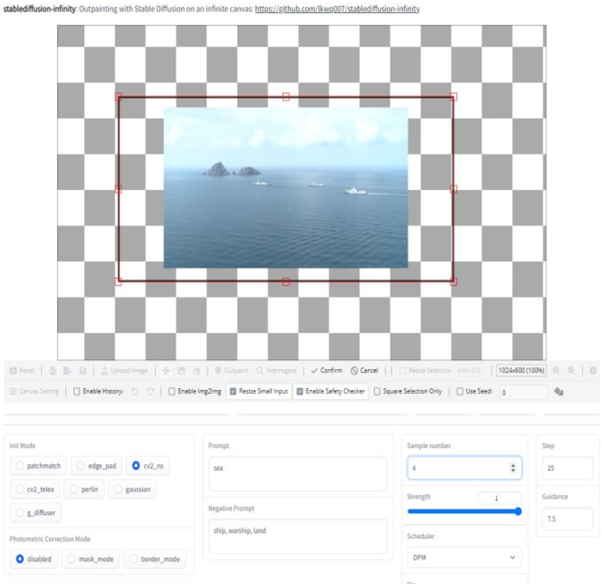


Fig. 1. The architecture of the Stable Diffusion model

Unlike the previous diffusion probability image generation model, which consumes resources exponentially as resolution increases, autoencoders are deployed in front and behind to insert/remove noise in a much smaller latent space rather than the entire image. Therefore, the main feature is that it can be used even with a general purpose graphics card by significantly reducing the resource consumption even when generating a relatively high resolution image.

III. Outpainting with Stable Diffusion

There are several open source projects for SD, including Stablediffusion-infinity (referred to as SD-I), a project that makes it easy to use outpainting with SD models through a web-based user interface [3]. Using this project, we can create a panoramic image of the desired size while preserving the properties of the existing image.


stablediffusion-infinity: Outpainting with Stable Diffusion on an infinite canvas: https://github.com/lkwq007/stablediffusion-infinity

1024x600 (100%)

□ Enable History: □ Enable Img2Img ☑ Resize Small Input ☑ Enable Safety Checker □ Square Selection Only □ Use Seed: 0

Init Mode: ○ patchmatch ○ edge_pad ● cv2_ns ○ cv2_telea ○ perlin ○ gaussian ○ g_diffuser

Photometric Correction Mode: ● disabled ○ mask_mode ○ border_mode

Prompt: sea

Negative Prompt: ship, warship, land

Sample number: 4

Strength: 1

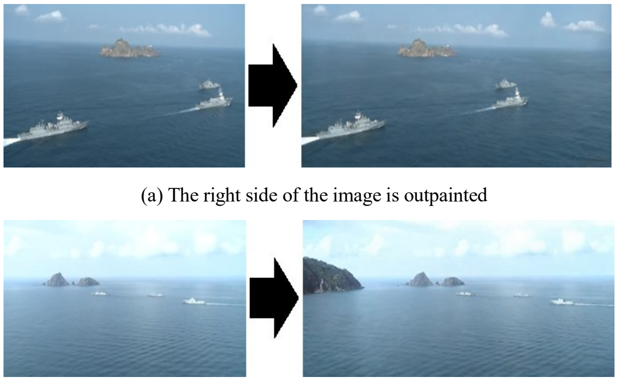
Scheduler: DPM

Step: 25

Guidance: 7.5

Fig. 2. Web UI of StableDiffusion-Infinity

We intend to create a panoramic view of the existing training data through SD-I and use it as training data for an object detector operating in smart glasses. For this purpose, we crawled open images of 640x480 size, which are similar to military CCTV images, and outpainted them in 900x412 size, which is the image size that can be obtained from smart glasses. The examples of the resulting images are shown in Figure 3.


(a) The right side of the image is outpainted

(b) The left side of the image is outpainted

Fig. 3. Example of SD-I generated images

As can be seen in the figure above, it was possible to create a panoramic image by expanding the desired part of the original image using SD-I, and it was ensured that the generated part matched well with the existing image without any awkwardness. In particular, more realistic and natural training data could be obtained by creating parts that are difficult to create naturally with existing generative technologies, such as clouds in Figure 3(a) or the island in Figure 3(b).

IV. Experiments

We prepared two types of datasets to verify the training validity of SD-generated panoramic images. First, 37 maritime surveillance images were collected at 900 x 412, the same resolution as the military smart glasses. In addition, 10 640 x 480 military maritime CCTV images were collected and training data was generated by outpainting in the form of a 900 x 412 panoramic view by SD-I.

A NVIDIA-A100 GPU was used for model training, and training was performed with the nano model of YOLOX[4] version 0.2.0 on ubuntu 20.04 lts. The hyperparameter settings used for training are as follows.

- Total epochs: 300
- Warmup epochs: 5
- No augmentation epochs: 15
- Basic learning rate per image: 0.01/64.0
- Minimum learning rate ratio: 0.05
- Weight decay factor: 5e-4

An NVIDIA-A100 GPU was used to verify model performance, and the rest of the experimental environment was the same as during model training.

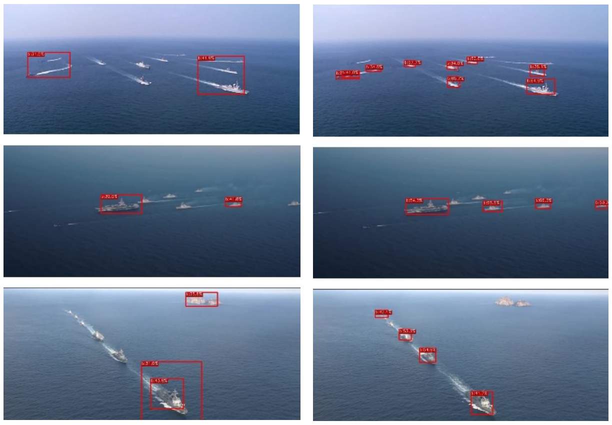
It was assumed that the training of the deep learning detection model that can be used for military smart glasses could only be done with few shot due to the lack of data, and the experiment was designed accordingly.

We first split 5 of the 37 original smart glasses datasets for training and 32 for verification, and added 10 outpainting data to training to see if performance improved.

TABLE I. COMPARISON OF DETECTION MODEL PERFORMANCE AFTER SD-I OUTPAINTED IMAGE ADDITION

	AP	AP-50	Recall
5 original only	6.11	24.60	22.89
5 original + 5 outpainting	12.59	43.61	27.46
5 original + 10 outpainting	**23.62**	**65.63**	**38.48**

As can be seen in the table above, both precision and recall increased to high levels when outpainting-based generated data was added compared to using only the original dataset. In particular, in the case of using the generated data of 10 shots, it was found that the AP-50 increased by about 2.5 times from 24.60% to 65.63% compared to the case of not using it, which had a positive effect on the training. The difference in performance was demonstrated in the following detection example.



(a) Original only

(b) Original + 10 outpainting

Fig. 4. Detection before and after adding SD-I generated images

The training data generated by SD-I and the training data generated by the traditional method were compared in terms of their effects on the training of a deep learning model. For this purpose, a deep learning model was trained by dividing the case of using 10 original CCTV images, the case of cropping according to the aspect ratio of the panoramic view, and the case of horizontally concatenating the same image. The results are shown in the table and graph below.

TABLE II. PERFORMANCE COMPARISION OF DETECTION MODEL WITH DATA GENERATED BY SD-I AND THE TRADITIONAL METHODS

	AP	AP-50	Recall
5 original + 10 original CCTV	18.80	58.60	35.33
5 original + 10 cropping	19.75	60.50	36.50
5 original + 10 concatenation	16.35	49.59	35.43
5 original + 10 outpainting (proposed)	**23.62**	**65.63**	**38.48**

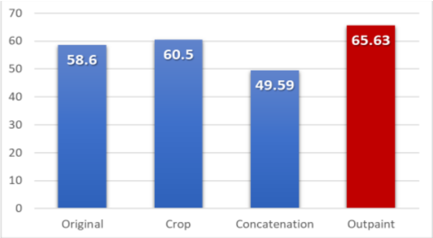


Fig. 5. Comparison of AP-50 results with training data

As can be seen from the experimental results, in all cases there was an improvement in performance compared to using only the original data. However, in the case of adding cropping data, there was a performance improvement of about 2% in AP-50 compared to the case of adding the original CCTV as training data, and in the case of adding concatenation data, there was a performance degradation of about 9%. On the other hand, a significant performance improvement of 7% was observed in the case of adding outpainting data. Therefore, it can be said that the proposed technique can be sufficiently used as training data for smart glasses.

V. CONCLUSION

In this paper, a data generation technique based on SD outpainting is presented as a method for generating training data for an object detection model to operate in military smart glasses, and the validity of the proposed technique is verified through experiments.

Various methods of processing existing data are used to supplement insufficient training data, and experiments confirm that a method using the excellent image generation capabilities of SD can also be a good solution.

In this paper, SD outpainting was used to generate panoramic view images for smart glasses, but the technique can also be used as a training data generation method to generate unusual unstructured image data, such as fish-eye images or ultra-wide-angle images.

ACKNOWLEDGMENT

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2023-00223530, Development of the artificial intelligence technology to enhance individual soldier surveillance capabilities)

REFERENCES

[1] Microsoft IVAS, [Online] Available : https://news.microsoft.com/source/features/digital-transformation/u-s-army-to-use-hololens-technology-in-high-tech-headsets-for-soldiers/

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models", arXiv:2112.10752 [cs.CV], 13 Apr 2022

[3] StableDiffusion-Infinity, [Online] Available : https://github.com/lkwq007/stablediffusion-infinity

[4] Z. Ge, S. Liu, and J. Sun, "YOLOX: Exceeding YOLO Seires in 2021," in Proc. CVPR. Jul. 2021.