Journal of KIIT. Vol. 23, No. 1, pp. 149-155, Jan. 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.1.149

A Diffusion-based Data Augmentation Method for SAR Object Detection

Daeyoung Han*[1], Yechan Kim*[2], Jonghyun Park*[3], Dongho Yoon*[4], and Moongu Jeon**

This work was supported by the Agency For Defense Development Grant funded by the Korean Government (UI220066WD)

Abstract

In recent times, diffusion models have garnered considerable attention as a key technology for generating images from text prompts. These models are especially recognized for their ability to produce high-resolution images with impressive accuracy. Another notable feature of diffusion models is their flexibility, as they can be easily fine-tuned through additional training to suit specific needs. This adaptability has made them useful in various tasks, one of which is data augmentation. In this study, we introduce a new data augmentation technique that enhances the training efficiency of object detection models. Specifically, our approach utilizes a diffusion model-based conditional generation method to create Synthetic Aperture Radar(SAR) images by taking input bounding boxes. This innovative approach demonstrates how diffusion models can effectively boost object detection performance.

요 약

최근 확산 모델은 텍스트 프롬프트로부터 이미지를 생성하는 핵심 기술로 상당한 주목을 받고 있다. 이러한 모델은 특히 높은 정확도로 고해상도 이미지를 생성하는 능력으로 인정받고 있다. 또 하나 주목할 만한 특징은 확산 모델의 유연성으로, 추가 학습을 통해 특정 요구 사항에 맞게 쉽게 미세 조정할 수 있다는 점이다. 이러한 적응력 덕분에 확산 모델은 다양한 분야에 적용되고 있으며, 그 중 하나가 데이터 증강이다. 본 연구에서는 객체 탐지 모델의 학습 효율성을 높이는 새로운 데이터 증강 기법을 제안한다. 구체적으로, 본 접근법은 바운딩 박스를 입력으로 받아 SAR(합성 개구 레이더) 영상을 생성하는 확산 모델 기반의 조건부 생성 기법을 활용한다. 이 혁신적인 방법은 확산 모델이 객체 탐지 성능을 효과적으로 향상시킬 수 있음을 보여준다.

Keywords

data augmentation, synthetic aperture radar, object detection, diffusion models

* PhD candidate in School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology
- ORCID[1]: https://orcid.org/0000-0003-0368-8675
- ORCID[2]: https://orcid.org/0000-0002-2438-3590
- ORCID[3]: https://orcid.org/0009-0005-5404-0707
- ORCID[4]: https://orcid.org/0009-0006-1514-7293
** Professor of School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology
- ORCID: https://orcid.org/0000-0002-2775-7789

· Received: Sep. 30, 2024, Revised: Nov. 20, 2024, Accepted: Nov. 23, 2024
· Corresponding Author: Moongu Jeon
School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology 123, Cheomdangwagi-ro, Buk-gu, Gwangju, 61005, Republic of Korea
Tel: +82-62-715-2406, Email: mgjeon@gist.ac.kr

Ⅰ. Introduction

Object detection technology plays a crucial role in various application areas such as autonomous driving, security surveillance, and medical image analysis. To maximize the performance of object detection models through machine learning, large-scale and high-quality training data is required. However, collecting such data often poses significant challenges in terms of time and cost. This issue is especially pronounced for Synthetic Aperture Radar(SAR) images, where obtaining diverse data requires substantial time and expense.

This study aims to generate SAR images utilizing the prior of Stable Diffusion and leverage these images to improve the performance of object detection models. Through this approach, it may be possible to overcome the data-hungry issue and contribute to the development of object detection models that perform well in a various domains. This paper first explains the theoretical background of diffusion models and proposes a data augmentation methodology based on pre-trained Stable Diffusion and ControlNet. The proposed methodology will then be experimentally validated, and the results will be analyzed to evaluate its effectiveness. Finally, the significance and limitations of this study, as well as directions for future research, will be discussed.

Ⅱ. Related Works

Data Augmentation(DA) has long been studied as a solution to the pervasive issue of data scarcity in the field of machine learning. It involves modifying existing data in various ways to create new data, thereby enhancing the model's generalization performance. In addition to the conventional approach, domain-specific generative models have been actively researched for DA recently. But there are few examples for dense prediction tasks including object detection. Bowles et al.[1] proposed a GAN training method to augment dataset for CT semantic segmentation. More recently, Bluethgen et al.[2] dealt with this topic by fine-tuning a pre-trained Stable Diffusion model with chest X-ray images. Despite the gap between natural images and their domain, it is verified that datasets augmented by synthetic medical images could help the performance for both tasks to improve.

However, these approaches are not appropriate to our task, SAR object detection. Because 80,000 and 377.110 images were required to achieve their reported performance respectively, which is a very challenging scale for SAR image domain. Hence we need to develop a novel DA method satisfying two condtions simultaneously: Data-efficiency and specialization for object detection task.

Ⅲ. Proposed Data Augmentation Method

3.1 Background

3.1.1 Denoising diffusion models

For a data x_0 drawn from the real data distribution $q(x)$, DDPM[3] defines the forward diffusion process that calculates noised samples $x_1,\cdots,x_T$ by adding Gaussian noise gradually to x_0 s follows.

$$q(x_t|x_0) = N\left(x_t; \sqrt{\overline{\alpha}_t}x_0, \sqrt{1-\overline{\alpha}_t}\right) \quad (1)$$

Here, $\overline{\alpha_0}=1$ and $\overline{\alpha_T}=0$. Hence it can be understood that x_T becomes complete Gaussian noise.

Now, if we know how to model the posterior distribution $q(x_{t-1}|x_t)$, we can get a real data sample x_0 by iterating the reverse step starting from a random Gaussian noise $x_T \sim N(0,I)$. However, there is no closed-form expression available to compute it unfortunately. Thus, we define a parameterized model $p_\theta(x_{t-1}|x_t)$, where the parameters θ are optimized using machine learning techniques so that $p_\theta(x_{t-1}|x_t)$ approximates $q(x_{t-1}|x_t)$.

To implement the single denoising step $p_\theta(x_{t-1}|x_t)$, we can define an neural network $\epsilon_\theta(x_t,t)$, which gets noisy data x_t and a timestep t as inputs. The parameters θ are optimized to minimize the following loss function:

$$L = |\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, t)|_2^2 \quad (2)$$

where $\epsilon \sim N(0,I)$ is the noise added to x_t.

3.1.2 Latent diffusion models

In contrast with the conventional diffusion models that undergo the diffusion process directly in RGB pixel space, latent diffusion models[4] perform this process in latent space compressed through an autoencoder. In other words, a data sample in pixel space $x \in R^{3\times H\times W}$ can be encoded into a 2-d latent vector $z = E(x) \in R^{C\times H/f\times W/f}$ via a pre-trained encoder E. The training and inference processes are conducted in the latent space instead of the raw data space. From a Gaussian noise vector $z_T \in R^{C\times H/f\times W/f}$, we can get z_0 by calculating the reverse diffusion process (3) and revert it into the raw data space using a pre-trained decoder D.

Latent diffusion models have the following advantages compared to the conventional diffusion models or other generative models for our task.

1. It significantly reduces the computational cost by performing the reverse diffusion process at a resolution that is f times smaller.
2. Stable Diffusion, a open-source latent diffusion model pre-trained with large-scale dataset, is available for fine-tuning with small-scale dataset.
3. It can be extended to various techniques developed based on the latent diffusion models, such as ControlNet.

3.1.3 Conditional data generation

Latent diffusion models are particularly specialized for conditional data generation, where they can generate desired data by inputting specific conditions. The most widely used example of a conditional generative model is Stable Diffusion, which can generate an image corresponding to an input natural language text prompt.

To input the text condition p into the model, a pre-trained natural language encoder τ such as CLIP[5] embeds the prompt into $c = \tau(p)$. Then the embedding vector c is injected into the cross-attention layers in the neural network ϵ_θ.

Besides the natural language prompt, the add-on model of Stable Diffusion called ControlNet[6] enable Stable Diffusion to operate more powerful conditional image generation from 2-d image conditions. This characteristic suits our task where the objective is synthesizing images given 2-d bounding box annotations.

3.2 Training method

The goal of this study is to develop a conditional generative model based on Stable Diffusion, which can generate SAR images based on given bounding box information. To achieve this goal, our latent diffusion model is first trained to generate SAR images randomly without receiving a bounding box as input. Then randomly-initialized ControlNet is attached to the model and the integrated model learns to generate SAR images corresponding to input bounding boxes.

3.2.1 Customizing stable diffusion

Practically, the scale of open SAR image datasets is far from enough to train a generative model from scratch. Therefore we focus on the customization techniques of the Stable Diffusion model that teach the pre-trained Stable Diffusion model a new concept about SAR images. However, Stable Diffusion is pre-trained with general pictures and illustrations on

RGB space, which is distant from the target domain of this study, SAR images on single-channel space. Thus the customization techniques such as DreamBooth[7] that inject a concept of specific objects and styles retaining the prior knowledge as possible with few images in the same domain do not fit into our case. Instead, we simply fine-tune all of the parameters in the Stable Diffusion model with SAR training data, as described in Fig. 1.

While the Stable Diffusion requires a natural language prompt for text-to-image generation, we don't need any textual condition to manipulate output images. So we fix the prompt embedding c in the inference process for stable generation. At this moment, we could set the natural language prompt as "An SAR image" and fix it during the training process. But we decide to randomly initialize the prompt embedding c_ϕ and allow it to be optimized along with the parameters θ, adopting the idea from Textual inversion[8].

3.2.2 Training ControlNet

As shown in Fig. 1, the ControlNet F_ψ is attached to the fine-tuned Stable Diffusion model and trained with data pairs of an SAR image and bounding boxes. Since it must receive conditions in the form of 2-d image, we conduct a pre-processing that draws the bounding boxes on the black background.

3.2.3 Dataset

We used the dataset called HRSID[9] as a training and test dataset. It consists of 5,604 cropped SAR images with 800×800 pixels containing ship objects, which can be divided into a training set (65%) and a test set (35%). Since the baseline model, Stable Diffusion 1.5 [10], is specialized for images with 512×512 size, we randomly crop the images and bounding boxes into 512×512 size during the training process.

3.3 Inferencing method

For efficient generation, we use the classifier-free guidance scale[11] $g = 2.5$ to generate a synthetic SAR image during 100 steps, which takes about 20 seconds with an NVIDIA RTX 3090 GPU. To evaluate the performance improvement, we synthesized 3,642 SAR images using bounding boxes of the 3,642 original images in the train set.

IV. Experimental Results

We evaluate our data augmentation method by measuring the performance improvement of an object detection model when trained with additional synthetic SAR data. In this study, we use the Mask R-CNN[12] model for evaluation. The detailed settings are set as summarized in Table 1.

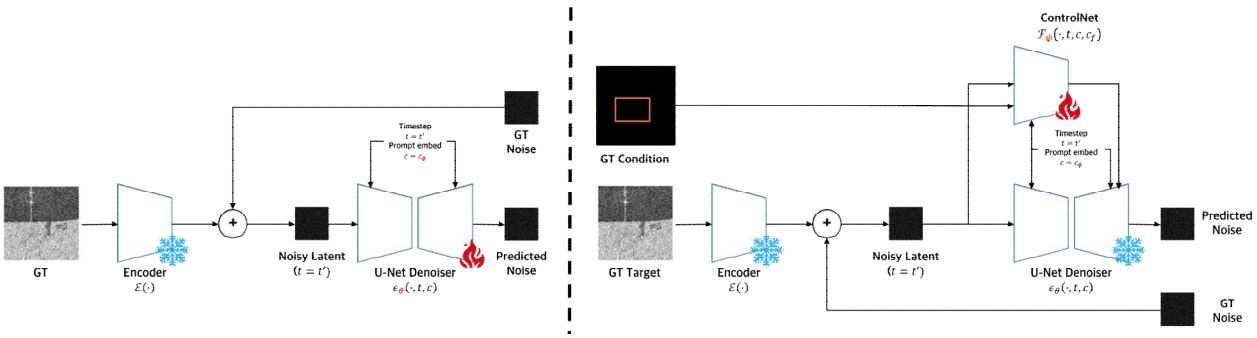

Fig. 1. Pipeline of Stable Diffusion fine-tuning for unconditional and conditional SAR image generation. Parameters in red are optimized during the fine-tuning process

Table 1. Hyperparameters for mask R-CNN training

Total epochs	Batch size	Learning rate	Optimizer
12	8	2.5e-3	SGD

4.1 Data generation

Fig. 2 presents the SAR images generated by the fine-tuned latent diffusion model and ControlNet. The first row shows bounding boxes sampled from the original dataset, while the subsequent rows display the generated SAR images conditioned on the corresponding bounding boxes, alongside the original data for comparison. The figure demonstrates that our model can generate SAR images far from the original sample while preserving the layout of ship objects. Notably, backgrounds featuring harbors and islands with complex shorelines are also intricately rendered.

One of the most notable characteristics of diffusion-based generative models is their inherent randomness. The final output varies depending on the initial random noise z_T and the random noises added at each denoising step, as illustrated in Fig. 2. This demonstrates that the training set can be infinitely augmented by repeatedly generating SAR images and ensures a diverse range of outputs, making the generative model highly effective for data augmentation purposes.

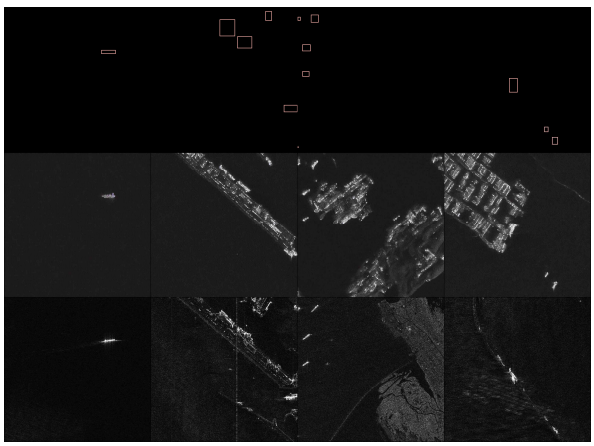

Fig. 2. Bounding boxes (1st row), corresponding synthetic SAR images generated by our generative model (2nd row), and the original SAR images from the training set (3rd row)

4.2 Qualitative results

This study aims to enhance an SAR object detection model by augmenting the training dataset with synthesized SAR images. The effectiveness of our data augmentation method can be evaluated by measuring how the object detection model's performance improves as the scale of the training data increases. As shown in Table 2, the performance of the ship object detection model improves as the ratio of additional synthesized SAR images to the original training data increases. Moreover, it outperforms a traditional method, horizontal flip, even with the least augmentation ratio of $r = 0.1$. Note that the horizontal flip doubles the amount of data effectively.

Table 2. Difference in mean AP metrics for object detection according to the ratio of augmented data to the original training data

Augmentation method	mAP	mAP50	mAP75
None	0.535	0.787	0.623
Horizontal flip	0.554	0.825	0.637
Ours ($r = 0.1$)	0.558	0.821	0.638
Ours ($r = 0.2$)	0.556	0.826	0.638
Ours ($r = 0.3$)	0.555	0.824	0.628
Ours ($r = 0.4$)	0.562	0.831	0.647
Ours ($r = 0.5$)	0.564	0.827	0.646

V. Conclusion

The scarcity of data is a fundamental challenge in machine learning tasks involving SAR images due to high costs and security concerns. This issue is even more pronounced in object detection tasks, which require precise object information for training. This study addresses the problem by synthesizing realistic SAR images incorporating bounding box information. Given the limited availability of training data for our generative model, we opted to fine-tune a pre-trained large-scale model rather than training from scratch. We selected Stable Diffusion, the most widely used open-source model, and proposed a method to custom-

ize it for our specific objectives. The experimental results demonstrate that the synthesized data produced by our model is of high quality and leads to measurable performance improvements in object detection models. To the best of our knowledge, this is the first data augmentation study for SAR object detection based on generative models.

References

[1] C. Bowles, et al., "Gan augmentation: Augmenting training data using generative adversarial networks", arXiv preprint, arXiv:1810.10863, Oct. 2018. https://doi.org/10.48550/arXiv.1810.10863.

[2] C. Bluethgen, et al., "A vision-language foundation model for the generation of realistic chest X-ray images", Nature Biomedical Engineering, Aug. 2024. https://doi.org/10.1038/s41551-024-01246-y.

[3] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models", Proc. of the 34th International Conference on Neural Information Processing Systems, Vancouver BC Canada, No. 574, pp. 6840-6851, Dec. 2020.

[4] R. Rombach, et al., "High-resolution image synthesis with latent diffusion models", Proc. of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, U.S., pp. 10684-10695, Jun. 2022.

[5] A. Radford, et al., "Learning transferable visual models from natural language supervision", International conference on machine learning, Virtual, pp. 8748-8763, Jul. 2021.

[6] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models", Proc. of the IEEE/CVF International Conference on Computer Vision, Paris, France, pp. 3836-3847, Oct. 2023.

[7] N. Ruiz, et al., "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation", IEEE/CVF conference on computer vision and pattern recognition., Vancouver, Canada, pp. 22500-22510, Jun. 2023.https://doi.org/10.1109/CVPR52729.2023.02155.

[8] R. Gal, et al., "An image is worth one word: Personalizing text-to-image generation using textual inversion", International Conference on Learning Representations, Kigali, Rwanda, May 2023.

[9] S. Wei, X.g Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation", IEEE Access, Vol. 8, pp. 120234-120254, Jun. 2020. https://doi.org/10.1109/ACCESS.2020.3005861.

[10] Hugging Face, "stable-diffusion-v1-5/stable-diffusion-v1-5", https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5 [accessed: Jun. 20, 2024]

[11] J. Ho and T. Salimans, "Classifier-free diffusion guidance", NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications, Dec. 2021.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn", IEEE international conference on computer vision, Venice, Italy, pp. 2961-2969, Oct. 2017. https://doi.org/10.1109/ICCV.2017.322.

Authors

Daeyoung Han

2019. 8 : BS degree, School of Electrical & Electronic Engineering, Yonsei University

Research interests: Generative AI, Image super-resolution

Yechan Kim

2020. 2 : BS degree, College of Education, Jeju National University

2021. 8 : MS degree, School of Electrical Engineering and Computer Science, GIST

Research interests: Visual understanding, Representation learning

Jonghyun Park

2021. 2 : BS degree, College of Information and Communication Technology Convergence, The University of Suwon

2023. 8 : MS degree, School of Electrical Engineering and Computer Science, GIST

Research interests: Computer vision, Machine learning

Dongho Yoon

2021. 2 : BS degree, Dept. of Mechanical System Design Engineering, Seoul National University of Science and Technology

Rsearch interests: Computer vision, Machine learning

Moongu Jeon

1999. 06 : MS degree, Dept. of Computer Science, University of Minnesota

2001. 06 : PhD degree, Dept. of Computer Science, Scientific Computation Program, University of Minnesota

2005. 9 ~ Present : Professor, School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology

Research interests: Artificial intelligence, Machine learning, Computer vision