Conditioning Latent Diffusion Model for Object Detection Dataset

Riku KAMADA* and Hiroshi TENMOTO**
* Advanced Course of Electronic and Information Systems Engineering
National Institute of Technology, Kushiro College
E-mail: s220707@kushiro.kosen-ac.jp
** Field of Information Engineering, Department of Creative Engineering
National Institute of Technology, Kushiro College
E-mail: tenmo@kushiro-ct.ac.jp

A dataset is called imbalanced if the number of samples belonging to each class is not equal and one class contains more samples than the others. A classifier using an imbalanced dataset can accurately classify classes with many samples but cannot learn enough data features to accurately classify classes with a small number of samples. Therefore, it has been reported that classification accuracy can be improved by using GANs and other generative models in classification tasks. In this paper, we propose a method for augmenting imbalanced object detection datasets by generating data and synthesizing images using a conditioned latent diffusion model. The proposed method uses semantic maps to condition the shape of the images to be generated. Then, based on the semantic map, the periphery of the generated image is removed. The resulting size of the image with the removed perimeter becomes the size of the annotation shape. Finally, the image is combined into a composable region of the background image and added to the dataset as annotated data. In experiments utilizing YOLOv8, the detection accuracy for classes with a limited number of samples improved by approximately 10% compared to the original dataset.

Key words: Imbalanced dataset, Object detection, Image generation

1. INTRODUCTION

In recent years, applications of machine learning for classification and detection in the field of computer vision have garnered attention. To achieve high accuracy in machine learning, a substantial amount of labeled data is required. However, real-world data often exhibits imbalances, and data collection can be challenging. Such datasets are referred to as "imbalanced," where there is an uneven distribution of samples among different classes. Existing research emphasizes the adverse impact of class imbalance on the performance of computer vision tasks [1]. Specifically, it is recognized that an imbalance in the number of samples belonging to each class can significantly degrade performance metrics. Particularly in the domain of real-world image datasets, class imbalance is a frequent challenge. This phenomenon arises from various factors, with a prominent cause being the specialized expertise required for data collection. For example, creating a medical image dataset requires not only specialized equipment but also the involvement of experts in the data collection process. Additionally, the manual effort involved in labeling datasets can worsen the issue of class imbalance. Many instances require extensive human annotation, contributing to an uneven distribution of samples across various classes. Furthermore, the inherent rarity of specific samples can naturally lead to class imbalance.

Therefore, generative models are gaining attention as a promising means to alleviate the challenges posed by class imbalance. These models possess the ability to generate synthetic data and address the rarity of specific samples, presenting a compelling approach to enhance the performance of computer vision systems facing imbalanced datasets. This paper specifically focuses on the utilization of generative models, with an emphasis on conditioned latent diffusion models, to explore the augmentation of imbalanced object detection datasets and improve overall detection accuracy.

2. GENERATIVE MODELS

Generative models, particularly in the realm of natural language generation, have gained widespread recognition, with ChatGPT being a prominent example [2]. However, our focus in this discussion shifts to the utilization of generative models in the domain of image generation. Generative models, leveraging sophisticated machine learning techniques, showcase the ability to produce highly realistic images by comprehending the intricacies of complex data. Notably, a recently introduced model known as "Stable Diffusion" [3] distinguishes itself for its proficiency in generating images based on textual instructions.

The core concept revolves around harnessing the capabilities of generative models to craft images that represent underrepresented classes, thereby contributing to the augmentation of datasets.

3. RELATED WORKS

In this section, we discuss common approaches for addressing insufficient data, namely, data augmentation, and existing data augmentation techniques employing generative models.

3.1 DATA AUGMENTATION

Data augmentation is a technique aimed at improving the performance of machine learning models by artificially expanding the training data through the manipulation of existing datasets. For instance, in the context of image data, augmentation involves operations such as rotation, flipping, scaling, cropping, and changes in color space. Widely used for its simplicity and effectiveness in implementation, certain operations such as rotation or flipping may have a detrimental impact on accuracy depending on the dataset.

3.2 GENERATIVE MODELS

In recent years, data augmentation techniques utilizing generative models have garnered attention. A generative model assumes a probability distribution for generating observed data and employs methods to estimate this distribution from the observed data. By generating images resembling the training data and incorporating them into the dataset, it is anticipated that the performance of the model can be enhanced. Particularly, data augmentation through Generative Adversarial Networks (GANs) is well-known to improve performance in classification tasks [4]. However, the application of GANs in object detection tasks is limited [5]. Datasets with insufficient data, or datasets exhibiting label imbalances like the one addressed in this study, pose a higher risk of overfitting with GANs. Achieving stable training under such conditions proves challenging.

Diffusion models exhibit more stable learning compared to GANs and can provide effective conditioning for tasks such as object detection. Our objective is to leverage the stability and conditioning advantages of diffusion models, specifically using Latent Diffusion Models, to address constraints associated with existing methodologies.

4. METHOD

Our approach involves training a Latent Diffusion Model (LDM) [6] conditioned on object detection labels. This allows us to generate images based on any given label, which are then seamlessly composited with backgrounds to create new data. The overall schematic of our system is illustrated in Fig.1.

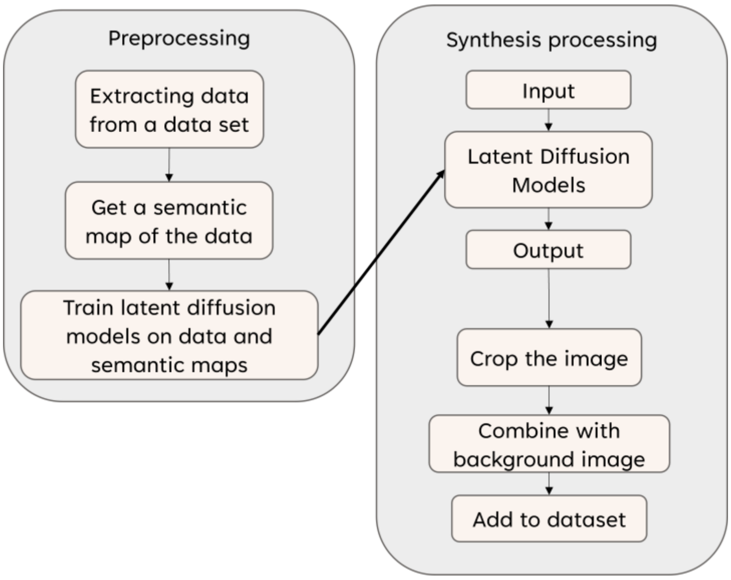
Preprocessing
Extracting data from a data set
Get a semantic map of the data
Train latent diffusion models on data and semantic maps
Synthesis processing
Input
Latent Diffusion Models
Output
Crop the image
Combine with background image
Add to dataset

Fig.1 Our system is designed to train the Latent Deformable Model (LDM) conditioned on semantic maps, enabling the generation of images based on arbitrary shapes.

Firstly, we discuss the preprocessing steps involved in training the LDM. Initially, the images we intend to generate are extracted from the dataset. Many object detection datasets contain multiple instances within a single image. To simplify the learning process, only the data of interest for training is cropped from the images. Next, semantic maps are obtained from the extracted images. This can be achieved either manually using tools like Labelme to create semantic maps, or automatically with tools like Deeplab. In this manner, the data and corresponding semantic maps required for training the LDM are prepared.

The following describes the synthetic process. Firstly, the shape of the generated image is determined as a semantic map. Initially, the semantic map is used as the input to generate an image using the Label-to-Image (LDM) model. Subsequently, the data corresponding to the object is cropped from the image generated according to the shape of the semantic map. By calculating the length and width of the cropped image at this point, it is possible to create labels for object detection. Finally, the cropped data is combined with a pre-prepared background image and added to the dataset, thereby augmenting the dataset with new data.

5. EXPERIMENTS

In the experiments, the detection accuracy using YOLOv8 was evaluated on a publicly available object detection dataset [7]. Due to the lowest proportion of the label "starfish" within all labels in this dataset, this study augmented the data for the "starfish" label and compared the original dataset with the proposed approach-generated dataset. The Label-to-Image (LDM) model was trained with a size of 128x128, and background images were selected from the training dataset. Fig.2 and Fig.3 depict images created by synthesizing images generated by LDM with background images.

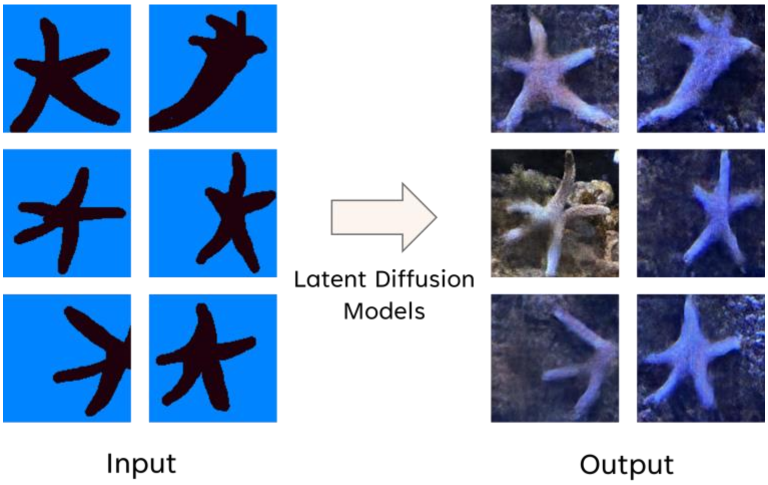

Fig.2 Output of the LDM subject to the semantic map. The ability to determine the shape improves the quality of the generated data and the generalization performance of the detection model.

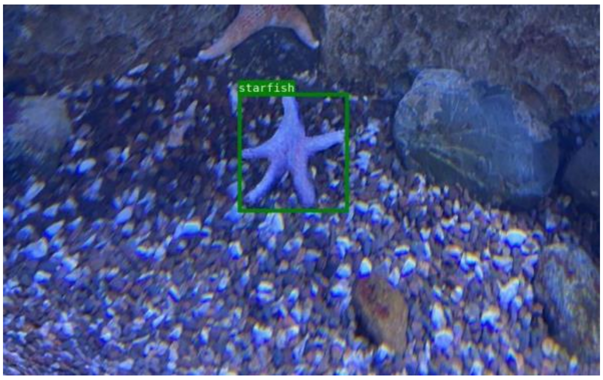

Fig.3 Image synthesized with background image using labelme. The size of the generated image is same as the size of the annotation.

The training of YOLOv8 was conducted using the default configuration parameters, and two types of mean Average Precision (mAP50, mAP50-95) were adopted as evaluation metrics.

Table 1 Experimental results for the original dataset and the dataset generated by our method.

	Original Dataset	**Ours**
mAP50	0.772	0.860
mAP50-95	0.555	0.673

Table 1 presents the experimental results. It is observed that the dataset generated by our method exhibits higher detection accuracy compared to the original dataset. Even for the stringent evaluation criterion of mAP50-95, our dataset surpasses the original dataset in terms of score.

6. CONCLUSION

In conclusion, our research demonstrated the effectiveness of the proposed method in enhancing datasets, particularly in the context of object detection tasks. Notably, the incorporation of conditioning has emerged as a pivotal factor in elevating the quality of the generated data. This nuanced strategy ensures that the generated images align more closely with the desired shapes specified in the input semantic maps, thereby enhancing the overall fidelity of the synthetic dataset.

As we look ahead, our future endeavors will involve extending our methodology's applicability by conducting comparisons with diverse datasets. This will enable a comprehensive assessment of its generalizability and performance across varied domains and scenarios. Additionally, a focused effort will be directed towards refining the composite process, aiming for further improvements in the seamless integration of generated content with background images.

References

[1] Lian Yu and Nengfeng Zhou, “Survey of Imbalanced Data Methodologies,” 2021, arXiv: 2104.02240.

[2] ChatGPT, November 2022. [Online]. Available: https://openai.com/chatgpt

[3] stablediffusion, August 2022 [Online]. Available: https://github.com/Stability-AI/stablediffusion

[4] Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha, “Data Augmentation Using GANs,” 2019, arXiv:1904.09135.

[5] H. Maeda, T. Kashiyama, Y. Sekimoto, T. Seto, H. Omata, “Generative adversarial network for road damage detection,” Computer-Aided Civil and Infrastructure Engineering., vol. 36, no. 1, pp. 47-60, 2021, doi: 10.1111/mice.12561.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Bjorn Ommer, “High-Resolution Image Synthesis With Latent Diffusion Models”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June, 2022, 10684-10695.

[7] Aquarium Dataset, Roboflow, November 2020. [Online]. Available: https://public.roboflow.com/object-detection/aquarium