2025 25th International Conference on Control, Automation and Systems (ICCAS)
November 4-7, 2025. Songdo Convensia, Incheon, Korea

SGGA: Semantic-Guided Generative Augmentation for Object Detection in Highly Imbalanced Disaster Imagery

Dayena Jeong[1], Dongwook Heo[2], Seonghyeok Ahn[1], Jonggeun Choi[1], and Sunglok Choi[2*]

[1]Dept. of Defence Applied Artificial Intelligence, Seoul National University of Science and Technology (SEOULTECH), Seoul, 01811, Korea ({pasteldiana, zboy20, choijon1}@seoultech.ac.kr)
[2]Dept. of Computer Science and Engineering, Seoul National University of Science and Technology (SEOULTECH), Seoul, 01811, Korea (hdwook3918@gmail.com, sunglok@seoultech.ac.kr) * Corresponding author

Abstract: Object detection in disaster images is often difficult because of class imbalance and a lack of labeled samples. To solve this problem, we introduce *Semantic-Guided Generative Augmentation* (SGGA), a new method that uses semantic masks to generate more samples for the rare classes. SGGA creates new images by changing clean road areas into Road-Blocked areas using mask-based sampling and prompt-guided inpainting, making sure the new objects appear in the right places. We filter the new images using CLIP similarity and LPIPS distance, ensuring high semantic and visual quality. Experiments on the RescueNet dataset show that SGGA improves Road-Blocked detection by +26.2% mAP@0.5 and +29.7% recall, beating other augmentation methods. Furthermore, t-SNE analysis confirms strong semantic alignment between real and SGGA-generated images. SGGA offers significant advantages, including spatial precision, contextual realism, and low annotation overhead, making it particularly suitable for practical deployment in disaster scenarios and other domains where spatial priors are available.

Keywords: Class imbalance, rare object detection, image data augmentation, generative data augmentation, stable diffusion, semantic-guided inpainting, disaster imagery

1. INTRODUCTION

Natural disasters threaten human lives, critical infrastructure, and ecosystems [1, 2], making robust disaster response systems critical for effective management [3–5]. However, disasters are inherently unpredictable, rare, and geographically localized [6], resulting in datasets with severe class imbalance. Figure 1 illustrates this challenge in the RescueNet dataset [7], where classes like Road-Blocked and Pool are significantly underrepresented compared to Vehicle.

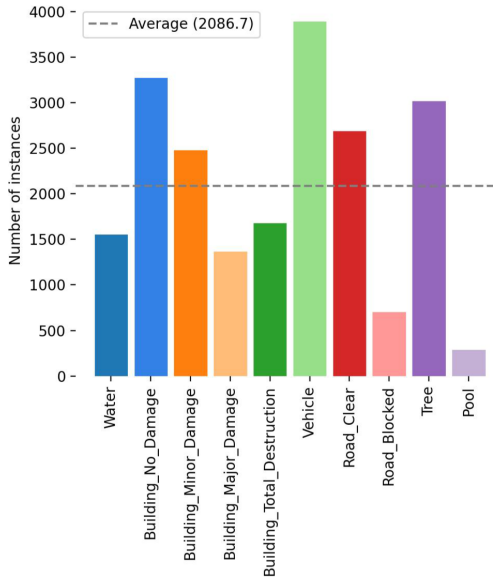

Fig. 1. The object class distribution in the RescueNet dataset [7]: Two classes (Road-Blocked and Pool) are especially rare, falling significantly below the average indicated by the gray dotted line.

This class imbalance severely impacts model performance, particularly for rare but critical classes. As shown in Table 2, our preliminary experiments with state-of-the-art object detectors [8] on RescueNet revealed extremely low accuracy for the `Road-Blocked` class (0.336 mAP@0.5). This poor performance stems from both the scarcity of training samples and the visual similarity between `Road-Blocked` and `Road-Clear` classes, which share global road context but differ only in local obstructive elements.

While conventional augmentation methods [9] and advanced generative approaches using GANs and diffusion models [10–13] have shown promise, they often lack precise spatial control needed for disaster imagery. Semantic-guided approaches like *SPADE* [14] and *GLIDE* [15] have introduced spatial conditioning, but their application to disaster scenarios with severe class imbalance remains underexplored.

To address these limitations, we propose *Semantic-Guided Generative Augmentation* (SGGA), a novel diffusion-based framework that leverages semantic masks to guide spatially controlled inpainting for minority class synthesis. Unlike conventional methods, SGGA ensures accurate object placement within semantically appropriate regions while maintaining contextual realism. Applied to the `Road-Blocked` class in RescueNet, SGGA achieves substantial improvements: +26.2% mAP@0.5 and +29.7% recall compared to baseline methods.

Our contributions include:

- A novel semantic-guided augmentation framework for disaster imagery.
- Integration of SGGA with YOLOv8 for real-time disaster detection.

- Extensive experiments demonstrating significant performance improvements for underrepresented classes.

2. SGGA: SEMANTIC-GUIDED GENERATIVE AUGMENTATION

SGGA is a four-stage framework designed to augment underrepresented classes in disaster datasets through mask-guided inpainting. The framework consists of the following stages:

1. **Source Image Selection**: Randomly select source images containing relevant semantic regions (e.g., `Road-Clear`) using conditional sampling to maintain data distribution.
2. **Semantic Mask Generation**: Extract and process binary masks from existing annotations, then apply random subregion sampling within target areas for localized inpainting control.
3. **Prompt-Guided Inpainting**: Apply Stable Diffusion inpainting with dual conditioning—spatial masks and disaster-specific text prompts—to synthesize realistic target instances (e.g., `Road-Blocked`).
4. **Image Quality Filtering**: Filter generated samples through multi-level validation, including semantic consistency (CLIP), visual fidelity (LPIPS/FID), and manual verification to ensure high-quality augmentation.

Figure 2 illustrates the complete SGGA pipeline, demonstrating how semantic guidance enables precise spatial control for minority class synthesis.

As shown in Figure 2, SGGA consists of four steps. At the source selection step, Road-Clear images were randomly selected from the dataset. At the mask generation step, semantic segmentation extracts road regions from annotations and samples random subregions as inpainting targets. When segmentation masks are included in the dataset, utilize them. At the inpainting step, we apply Stable Diffusion [16] with binary masks and disaster-specific prompts. For example, it is as follows.

- "an aerial view of a destroyed road and buildings"
- "an aerial view of a destroyed road and cars"
- "a view of a building and a road that has been destroyed"
- "an aerial view of houses and a road that have been destroyed"
- "an aerial view of houses and a road that were destroyed by hurricane florence"

At the quality filtering step, samples are retained that pass semantic consistency (CLIP > 0.80), visual fidelity (LPIPS < 0.68), and manual verification thresholds.

The key innovation of SGGA lies in its dual conditioning approach, where semantic masks provide precise spatial control while text prompts guide content generation. We employ carefully designed disaster-specific prompts such as *"a road buried under collapsed buildings"* and *"a road blocked by fallen trees and debris"* to ensure realistic obstruction synthesis. Random subregion sampling within semantic masks enhances diversity while preserving global scene context and preventing interference with surrounding objects.

To guarantee high-quality synthetic data, SGGA incorporates comprehensive quality filtering using multiple metrics. Semantic consistency is validated through CLIP similarity scores, visual fidelity is assessed via LPIPS distances, and distributional alignment is measured using FID scores. This rigorous validation process ensures that only semantically coherent and visually realistic samples are retained for training, effectively addressing class imbalance while maintaining data quality.

3. EXPERIMENTS

3.1 Dataset and Experimental Setup

Experiments were conducted on the RescueNet dataset, converted to object detection format. The dataset contains 4,494 UAV images with significant class imbalance, particularly for the `Road-Blocked` class. YOLOv8n was used as the detection model, trained on augmented datasets with various augmentation strategies.

3.2 Results and Analysis

Table 1. Quantitative comparison of image generation quality across methods. (CLIP: semantic alignment, LPIPS: perceptual distance, FID: distributional gap)

Methods	CLIP (↑)	LPIPS (↓)	FID (↓)
Diffusion with LoRA [17, 18]	0.865	0.666	1.284
SGGA (Ours)	0.957 (+10.6%)	0.641 (-3.8%)	1.043 (-18.8%)

Synthetic Image Quality: Table 1 demonstrates the enhanced image quality achieved by SGGA compared to Diffusion LoRA, showing significant improvements across CLIP, LPIPS, and FID metrics. SGGA outperformed Diffusion LoRA in both semantic and perceptual quality metrics, achieving a CLIP similarity of 0.957 (+10.6%) and an LPIPS score of 0.641 (-3.8%). t-SNE visualizations confirmed that SGGA-generated embeddings closely align with real disaster images. Figure 3 presents a t-SNE visualization of CLIP embeddings, demonstrating that SGGA-generated samples form tight clusters near real `Road-Blocked` instances.

As shown in Figure 4, SGGA generates more realistic and contextually aligned disaster imagery compared to Diffusion LoRA, which often introduces structural inconsistencies.

Object Detection Performance: Table 2 summarizes the detection performance for underrepresented classes using various augmentation techniques. SGGA achieved a 26.2% improvement in mAP@0.5 for the `Road-Blocked` class compared to the baseline, demonstrating its effectiveness in addressing class imbalance. As shown in Table 3, overall model performance across all classes also improved, with SGGA achieving the highest mAP@0.5 of 0.737.

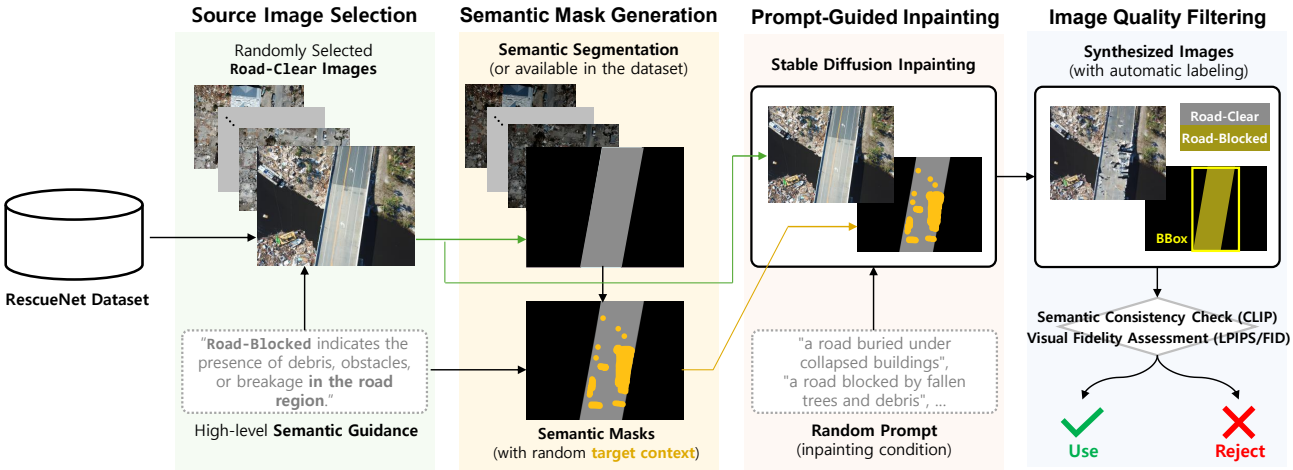

Fig. 2. Semantic-Guided Generative Augmentation (SGGA) pipeline: At the first step, select random source images. At the second step, generate semantic masks including random target masks. At the third step, synthesize new target images by inpainting the target masks using stable diffusion and random prompts. At the final step, semantic consistency and visual fidelity are evaluated to select high-quality images.

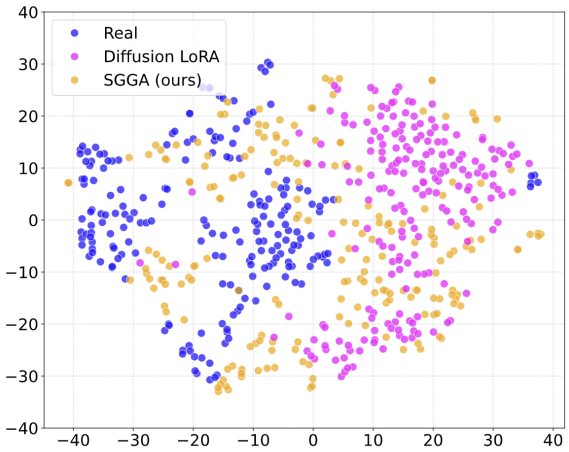

Fig. 3. t-SNE visualization of CLIP embeddings for real and generated `Road-Blocked` images. SGGA-generated samples align closely with real images, indicating strong semantic consistency.

Table 2. Class-specific detection performance metrics for the Road-Blocked class. (Pr: Precision, Re: Recall, mAP: mean Average Precision at IoU thresholds 0.5 and 0.5:0.95.)

Methods	**Pr** (↑)	**Re** (↑)	**mAP** (↑) **@0.5 / 0.5:0.95**
No Augmentation	0.448	0.290	0.336 / 0.194
Conventional Aug. [8]	0.491	0.344	0.351 / 0.197
Diffusion with LoRA [17, 18]	0.573	0.333	0.373 / 0.223
SGGA (Ours)	0.568	0.376	0.424 / 0.240

4. CONCLUSION

We introduced SGGA, a semantic-guided augmentation framework that addresses class imbalance in disaster object detection. By leveraging segmentation masks and diffusion inpainting, SGGA generates high-quality synthetic instances that improve detection performance for underrepresented classes. Experiments on the RescueNet dataset demonstrated significant improvements in both image quality and detection metrics. Future work will aim to extend SGGA to other domains with imbalanced classes, reduce reliance on pixel-level annotations, and optimize the combinations of real and synthetic data for robust deployment. Additionally, while the current experiments are conducted exclusively on UAV-based disaster imagery datasets such as RescueNet, future research will focus on generalizing SGGA to encompass other perspectives, including first-person view images.

Table 3. Overall detection performance metrics across all classes using different augmentation strategies.

Methods	**mAP** (↑) **@0.5**	**mAP** (↑) **@0.5:0.95**
No Augmentation	0.725	0.532
Conventional Aug. [8]	0.719	0.518
Diffusion with LoRA [17, 18]	0.733	0.534
SGGA (Ours)	0.737	0.532

REFERENCES

[1] S. Banholzer, J. Kossin, and S. Donner, *The Impact of Climate Change on Natural Disasters*. Springer Netherlands, 2014.

[2] S. Hallegatte, *Shock Waves: Managing the Impacts of Climate Change on Poverty*. World Bank Publications, 2016.

[3] X. Zhu, J. Liang, and A. Hauptmann, "MSNet: A Multilevel Instance Segmentation Network for Natural Disaster Damage Assessment in Aerial Videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[4] Y. Pi, N. D. Nath, and A. H. Behzadan, "Convolutional neural networks for object detection in aerial imagery for disaster response and recovery," *Advanced Engineering Informatics*, vol. 43, 2020.

Fig. 4. Qualitative comparison of `Road-Blocked` image generation. SGGA preserves spatial coherence and generates contextually realistic obstructions compared to Diffusion LoRA.

[5] V. Spasev, I. Dimitrovski, I. Chorbev, and I. Kitanovski, "Semantic Segmentation of Unmanned Aerial Vehicle Remote Sensing Images Using SegFormer," in *Proceedings of the International Conference on Intelligent Systems and Pattern Recognition*, 2024.

[6] M. Dilley, *Natural Disaster Hotspots: A Global Risk Analysis*. World Bank Publications, 2005, vol. 5.

[7] M. Rahnemoonfar, T. Chowdhury, and R. Murphy, "RescueNet: A High Resolution UAV Semantic Segmentation Dataset for Natural Disaster Damage Assessment," *Scientific data*, vol. 10, no. 1, 2023.

[8] G. Jocher, A. Chaurasia, J. Qiu, A. Stoken, J. Borovec, A. Kharbat *et al.*, "Ultralytics YOLOv8 Documentation," 2023, https://docs.ultralytics.com.

[9] C. Shorten and T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, 2019.

[10] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using GAN for improved liver lesion classification," in *Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018.

[11] S. Milz, T. Rudiger, and S. Suss, "Aerial GANeration: Towards Realistic Data Augmentation Using Conditional GANs," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[12] A. Toker, M. Eisenberger, D. Cremers, and L. Leal-Taixé, "SatSynth: Augmenting Image-Mask Pairs through Diffusion Models for Aerial Semantic Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[13] S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, and D. J. Fleet, "Synthetic Data from Diffusion Models Improves ImageNet Classification," *arXiv preprint arXiv:2304.08466*, 2023.

[14] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic Image Synthesis with Spatially-Adaptive Normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," *arXiv preprint arXiv:2112.10741*, 2021.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, 2021.

[18] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.