IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 October 2025, accepted 28 October 2025, date of publication 4 November 2025, date of current version 11 November 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3629349

RESEARCH ARTICLE

Augmenting Crosswalk Segmentation With Diffusion-Generated Data for Adverse Weather and Lighting Conditions

KREŠIMIR ROMIĆ, HRVOJE LEVENTIĆ, MARIJA HABIJAN, AND IRENA GALIĆ, (Member, IEEE)

Faculty of Electrical Engineering, Computer Science and Information Technology, 31000 Osijek, Croatia

Corresponding author: Krešimir Romić (kresimir.romic@ferit.hr)

This work was supported in part by Croatian Science Foundation under Project IP-2024-05-9492.

ABSTRACT Training deep learning models for semantic segmentation requires substantial annotated datasets, which are often limited in diversity and size. This study investigates the effectiveness of using diffusion-generated synthetic data to enhance crosswalk segmentation models for assistive navigation systems serving visually impaired individuals. We hypothesized that a Stable Diffusion model fine-tuned on a small collection of 150 first-person view (FPV) images, predominantly captured in sunny conditions, could generate valuable training data representing diverse weather scenarios not present in the original dataset. To test this hypothesis, we created 1500 synthetic images with varying weather conditions using our fine-tuned Stable Diffusion model. We then trained several U-Net-based semantic segmentation models on different combinations of synthetic and original images. For evaluation, we developed a new high-quality test dataset containing 300 annotated real-world images captured across diverse weather conditions. Our results confirmed that models trained on a combination of real and synthetic images significantly outperformed those trained exclusively on the limited real-world dataset. The mixed-data model demonstrated superior generalization to weather conditions absent from the original training data, achieving higher performance metrics across all test scenarios. These findings underscore the potential of diffusion models for data augmentation when authentic datasets are constrained. The pipeline we propose offers an efficient, scalable, and adaptable approach applicable to other domains requiring synthetic data enhancement.

INDEX TERMS Crosswalk segmentation, data augmentation, stable diffusion, visually impaired.

I. INTRODUCTION

Machine learning has become standard, and it is widely adopted in modern computer vision tasks, offering impressive results in various applications. However, machine learning methods often require large amounts of training data to achieve robust performance. Although a certain number of public datasets are available for general-purpose applications such as object detection or facial recognition, this is not always the case for specific domains. One such example is the segmentation of pedestrian crosswalks, particularly for use in wearable camera systems for visually impaired people, where crosswalk is observed from the pedestrian point of view.

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

Segmentation of crosswalks in real-world environments poses some challenges due to a specific first-person view (FPV) approach and varying weather and lighting conditions, including rain, bright sunlight, clouds, or night scenarios. Publicly available datasets for this specific task are rare and limited, making it difficult to train and deploy models that are accurate and robust in such diverse conditions. Without such robustness, models may fail in critical real-world situations, particularly in safety-critical applications like assisting visually impaired pedestrians or integrating into wearable navigation systems.

To address this limitation, we propose a novel approach that uses diffusion models to generate synthetic datasets for training the crosswalk segmentation models. Diffusion models, such as Stable Diffusion, are capable of generating

images simulating various environmental conditions through text-guided image generation. Fine-tuning a diffusion model on a smaller, limited dataset teaches the model to conform to the pedestrian perspective (FPV) when generating images, while textual prompts instruct the model to generate images with varied weather and lighting conditions. The combination of textual prompting for weather conditions and fine-tuning for pedestrian perspective adherence has enabled the generation of a large synthetic dataset, with images in diverse weather and lighting conditions, that can be used to train robust semantic segmentation models. Using this dataset, we train a U-Net model and validate its performance on a separate high-quality test dataset, demonstrating that the inclusion of diverse synthetic data significantly improves segmentation performance compared to training on a smaller, limited dataset, particularly under aggravating conditions. The main goal of this work is to demonstrate that a small, limited dataset can be effectively augmented with synthetically generated images and that such augmentation can significantly improve the downstream trained models' performance.

The key contributions of this study are as follows:

1) **Proposal of a fine-tuned diffusion model for the generation of synthetic crosswalk datasets** – Using a limited dataset, we fine-tuned a Stable Diffusion model to generate a large dataset of high-fidelity first-person view synthetic crosswalk images, ensuring realistic representation of pedestrian perspectives.
2) **Enhancing model robustness through diverse weather and lighting conditions** – By introducing controlled variations in synthetic data, including adverse weather and nighttime scenarios, we significantly improved the generalization capabilities of crosswalk segmentation models.
3) **Comprehensive validation demonstrating superior segmentation performance** – Through rigorous evaluation on a diverse real-world test set, our approach achieved substantial accuracy gains, outperforming models trained exclusively on limited real-world data.

This study not only addresses the lack of publicly available datasets for crosswalk segmentation but also presents a scalable pipeline for creating augmented datasets for other computer vision tasks. The results highlight the potential of diffusion models in addressing domain-specific dataset limitations and improving robustness in real-world applications.

The remainder of this paper is organized as follows: Section II reviews related work on synthetic data generation, diffusion models, and crosswalk segmentation. Section III describes the proposed methodology, including data collection, synthetic dataset generation, and model training. Section IV presents experimental results and a discussion of the findings. Finally, Section V concludes the paper.

II. RELATED WORK

The research conducted in this paper encompasses three main fields: diffusion models for synthetic image generation, machine learning-based segmentation of objects (e.g. crosswalks), and computer vision-based assistance systems for people with visual impairment.

Diffusion models are a class of generative models that have gained significant attention in recent years for their ability to produce high-quality images. These models work by progressively transforming noise into a structured image through an iterative denoising process [1]. By training on large datasets, these models generate highly detailed and diverse images, making them a strong competitor to Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) in tasks such as image synthesis or inpainting [2]. Stable Diffusion, developed by Stability AI, is a state-of-the-art open-source model for high-resolution image generation from text prompts [3]. As a Latent Diffusion Model (LDM), it operates in a compressed latent space rather than directly in pixel space, significantly reducing memory consumption while preserving high image fidelity. Authors in [4] suggest the use of stable diffusion model to generate diverse synthetic data in order to improve ML model's ability to adapt to unknown new data. Similarly, stable diffusion is used as an effective data augmentation technique in [5] and [6].

When compared to GANs and VAEs, diffusion models achieve competitive or superior visual quality in image synthesis [7]. They offer more stable training and, crucially, robust text-to-image conditioning mechanisms [8] with high semantic alignment [9]. This capability is especially important for our application, as it allows controlled generation of diverse weather and lighting conditions through simple prompt modifiers. In contrast, GANs and VAEs are less flexible for prompt-driven variation and often require complex conditioning schemes. For this reason, we adopt a diffusion-based augmentation pipeline in this study.

Machine learning techniques are often utilized for crosswalk segmentation task, but mostly in the context of autonomous driving like in [10], [11], and [12]. In [13], authors deal with crosswalk segmentation in the context of assistive systems for visually impaired, and they propose U-net based model for segmentation. As noted in [14], the U-Net architecture is widely used for semantic segmentation due to its strong performance and adaptability across different domains. Tian et al. propose a novel system for crosswalk scenes, which additionally detects other objects beside crosswalks, such as vehicles, pedestrians, and traffic light status [15]. Review [16] focuses on the challenges and solutions for semantic segmentation in adverse weather and lighting conditions, relevant to enhancing model robustness in real-world scenarios. On the other side, authors in [17] deal with detecting wear and tear on crosswalks due to traffic and age, and that is also an important information to take into

(a) Real-world images from initial dataset

(b) Stable diffusion generated images (before fine-tuning)

(c) Stable diffusion generated images (after fine-tuning)

FIGURE 1. **Examples of real-world and generated images used for this approach.**

consideration when generating synthetic data that should look realistic.

Surveys on assistive technologies for visually impaired [18], [19], [20] often show increasing use of well-known machine learning techniques. However, since most of these works focus on autonomous driving and similar applications, there is still room for improvements in this field, especially for specific tasks like FPV crosswalk detection and segmentation.

III. METHODOLOGY

This section outlines the methodology of the proposed approach. It begins with the synthetic data generation process (Subsection III-A), followed by the annotation procedure (Subsection III-B). Finally, Subsection III-C details the training process and preparation for result validation.

A. GENERATING SYNTHETIC DATA

The first objective was to expand the initial limited dataset, which contained 150 real-world images of crosswalks. This initial dataset was a result of our previous work [21] in crosswalk detection. The images were captured from a first-person perspective, with a wearable chest-mounted camera, simulating a pedestrian's view while approaching a crosswalk (Fig. 1a). However, all images were collected exclusively during daylight hours under predominantly sunny conditions. Given the dataset's small size and lack of diversity in weather and lighting conditions, training deep learning models on this data alone raised concerns regarding generalization and robustness. To avoid these limitations, we used diffusion models to generate additional synthetic images that closely match the visual characteristics and style of the original dataset, while improving the diversity of lighting and weather conditions, thus effectively augmenting the training data.

TABLE 1. **Fine-tuning parameters for Stable Diffusion.**

Parameter	Value / Description
Pretrained Model	Stable Diffusion v1-4
Fine-Tuning Data	150 real-world crosswalk images (first-person perspective)
Image Resolution	512×512 pixels
Batch Size	4
Epochs	10
Optimizer	AdamW
Learning Rate	1×10^{-5}
Loss Function	Mean Squared Error (MSE)
Noise Injection	Gaussian noise added to latents
Train. Framework	PyTorch

TABLE 2. **Prompts for generating synthetic images and number of generated images.**

Prompt	No. of images
"a crosswalk image"	1500
"a crosswalk image, sunny weather"	600
"a crosswalk image, cloudy weather"	300
"a crosswalk image, rainy weather"	300
"a crosswalk image, night conditions"	300
Total	3000

Stable Diffusion is one of the most widely used implementations of diffusion models for image generation. It is an open-source model designed to generate high-resolution images from text prompts. However, images generated directly with generic prompts do not match the style of the initial dataset images and the generated images vary

(a) Sunny weather

(b) Cloudy weather

(c) Rainy weather

(d) Night conditions

FIGURE 2. Examples of stable diffusion generated images using additional condition prompts.

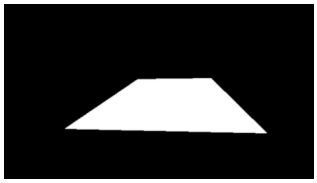
significantly from the pedestrian perspective, as visible in Fig. 1b. Therefore, a fine-tuning process was conducted in order to teach the model to understand and reproduce the specific characteristics of initial crosswalk images. In fine-tuning, model learns the unique visual patterns like angle and perspective and finally reduces the variability in generated images. Fine-tuning also creates a tighter mapping between the desired prompt and visual characteristics of the initial dataset. Overall, this process helps ensure that synthetic data maintain the essential characteristics of real crosswalk images while enabling large-scale data generation. The full set of fine-tuning parameters is summarized in Table 1 and the fine-tuned diffusion model was made publicly available under Apache 2.0 license on Hugging Face [22]. Image generation is done by asking the model to generate image with the base prompt: ''a crosswalk image''. Examples of images generated with fine-tuned stable diffusion model are shown in Fig. 1c.

Since Stable Diffusion is a text-to-image model, the base prompt can be further modified to generate images under different weather and lighting conditions. The original dataset primarily consists of images captured in dry weather and daylight conditions, which means that synthetic images generated with the base prompt alone will replicate these characteristics. However, by incorporating additional prompt modifiers (e.g., ''rainy weather''), the model can generate crosswalk images with the same perspective but under different environmental conditions. For this study, four key environmental conditions (sunny, cloudy, rainy, and night) were introduced as modifications to the base prompt to ensure greater diversity in the dataset. A total of 3,000 synthetic images were generated as detailed in Table 2. Examples of synthetic crosswalk images generated under adverse weather and lighting conditions are shown in Fig. 1c.

B. ANNOTATION PROCESS

To achieve accurate semantic segmentation of crosswalk regions, a precise annotation process was conducted for the newly generated synthetic data. The annotation was performed manually by marking the crosswalk area, which is typically represented as an irregular quadrilateral polygon. To streamline this process, a custom annotation interface was developed, allowing users to define the crosswalk region by clicking on its four corner points (Fig. 3a). For each annotated image, a corresponding binary mask file was generated. The mask image consists of two distinct regions: the crosswalk area, represented in white (foreground), and the background, represented in black (Fig. 3b). This binary segmentation serves as ground truth for training the segmentation model.

(a) Marking annotation polygon



(b) Binary mask

FIGURE 3. Annotation process.

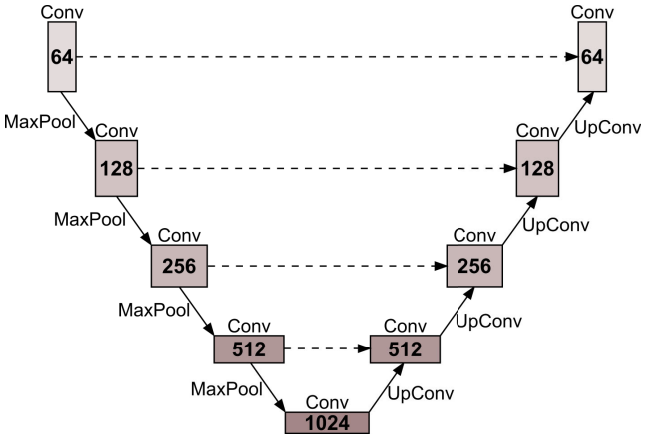


FIGURE 4. U-Net architecture used in this approach.

Additionally, the same annotation procedure was applied to 300 real-world images, which were later used for validating the segmentation model's performance. The developed annotation interface enables an efficient workflow, allowing the user to systematically process all images within a given dataset folder.

C. TRAINING U-NET MODEL

To perform the crosswalk semantic segmentation, modified U-net architecture was implemented (Fig. 4). The implemented U-Net architecture follows a symmetric encoder-decoder structure designed to process and segment images effectively. The encoder path consists of four sequential blocks, each performing double convolution operations with ReLU activation, progressively increasing the feature depth from 64 to 512 channels while reducing spatial dimensions through max pooling operations. At the network's center, a bottleneck layer with 1024 channels captures deep feature representations at the lowest resolution. The decoder path mirrors the encoder with four upsampling blocks that gradually restore spatial dimensions while reducing the channel depth back to 64 through transposed convolutions and ReLU activations. The network ends with a final 1×1 convolution layer followed by a sigmoid activation, producing binary segmentation outputs suitable for crosswalk segmentation. This design enables an effective capture of both fine-grained spatial details and high-level semantic information, making it well-suited for precise crosswalk segmentation in various image conditions.

The experiments performed in this study involved training multiple models on different datasets, each consisting of annotated images paired with corresponding segmentation masks. The following training configuration was used for each training run: The dataset was split with 80% allocated to training and 20% to validation. Images were preprocessed to a consistent size of 256×256 pixels. The model was trained using a batch size of 16, allowing for efficient gradient computation and memory utilization. The training process spanned 50 epochs with early stopping implemented. An Adam optimizer was utilized with a learning rate of 1×10^{-3} for balance between convergence speed and model stability. Model optimization was guided by a combined loss function defined as the sum of Binary Cross-Entropy and Dice loss, balancing pixel-wise accuracy and region overlap. During inference, a threshold of 0.5 was applied to the sigmoid output to obtain the final binary segmentation mask. This configuration aimed to optimize the U-Net architecture's performance in segmenting crosswalks from images.

To comprehensively assess the impact of synthetic data on model performance, we trained four separate models using the same U-Net-based architecture but with varying training datasets. The first model was trained exclusively on the initial, limited dataset of 150 real-world crosswalk images. The second model utilized 1500 synthetic images generated by a fine-tuned Stable Diffusion model using a simple base prompt (''a crosswalk image''). The third model was trained on another set of 1500 synthetic images, but these were generated using an expanded prompt that incorporated diverse weather and lighting conditions. Finally, the fourth model was trained on a mixed dataset comprising both the 150 real-world images and the 1500 synthetic images obtained using the expanded prompt, resulting in a total of 1650 training images. Each of these models was subsequently evaluated on a high-quality real-world test set (300 images) to determine their generalization ability and performance under realistic conditions. Synthetic and real-world datasets used in this evaluation have been made publicly available and described in [23].

IV. RESULTS AND DISCUSSION

To evaluate the trained models and obtain relevant performance metrics, we created a high-quality real-world test dataset by collecting and annotating images captured under diverse environmental conditions. First, we extracted still images (frames) from video sequences recorded while a person wearing a chest-mounted camera approached a pedestrian crosswalk. Using video frames instead of static images helps capture realistic noise and motion artifacts, better reflecting real-world scenarios [24]. From the extracted

images, we selected 300 images and collected them into four subsets (environmental scenarios): sunny (120 images), cloudy (60 images), rainy (60 images) and night (60 images). Afterwards, we manually annotated the images (as explained in section III-B) to obtain ground-truth masks. Although this study relied on manual annotation of images to guarantee accurate ground truth, future work could explore automating the generation of both images and corresponding segmentation masks. Such approaches may substantially reduce annotation effort and improve scalability [25]. Importantly, the manual annotations provided here establish a solid benchmark against which the reliability of automatic annotation methods can later be assessed.

For each trained model, evaluation on the test dataset was performed using a standard evaluation procedure: process each image from the dataset with the trained model (inference), store the predicted segmentation mask and compare it to the ground truth mask. The results are presented with the following metrics that illustrate the relationship between the obtained segmentation predictions and the ground truth masks:

- Dice Coefficient – measures the overlap between predicted and ground truth segmentation masks, calculating the harmonic mean of precision and recall
- Intersection over Union (IoU) – compares the intersection and union of predicted and actual regions
- Pixel Accuracy – represents the percentage of correctly classified pixels across the entire image
- Precision – indicates the percentage of predicted pixels that are actually crosswalk
- Recall – measures the percentage of actual crosswalk pixels successfully detected by the model.

All metrics were calculated as average values for the entire set of test images. Additionally, we also show the performance on each of the test scenarios (sunny, cloudy, rainy and night). Table 3 presents the segmentation performance of the four trained models across different test scenarios, highlighting the improvements achieved through synthetic data augmentation. The results demonstrate that incorporating diffusion-generated synthetic data significantly enhances model performance, particularly in handling diverse weather and lighting conditions.

It is important to note that the high-quality test dataset (300 images) explained here is used solely for the evaluation of the trained models and never used in training, except for the cross-validation training run to obtain the benchmark performance on a high-quality dataset. The goal of this benchmark training run is to illustrate how close can the performance on a synthetically augmented dataset approach the performance on a high-quality gold standard dataset.

The baseline model, trained exclusively on an initial, limited dataset of 150 real-world images, achieved an overall Dice coefficient of 0.6899. While reasonable, this performance reflects the dataset's limitations, as it primarily consists of images captured in sunny conditions. As a result, the model struggles to generalize to adverse conditions, particularly night scenarios, where the Dice score drops to 0.1878. This highlights the challenges of training deep learning models with small, biased datasets that lack environmental diversity.

The second model, trained solely on 1500 synthetic images generated using a simple base prompt (without targeted prompts for weather and lighting conditions), performed the worst overall, with a Dice coefficient of 0.5954. The model particularly struggled in nighttime conditions (Dice = 0.0670) and showed lower generalization across all test scenarios. This suggests that solely using the synthetic data for training, despite much larger number of samples, is insufficient unless the generated images capture relevant real-world weather and lighting variations.

A significant improvement was observed with the third model, trained on 1500 synthetic images generated using an expanded prompt incorporating diverse weather and lighting conditions. This model, despite being trained solely on synthetic images, achieved a Dice coefficient of 0.8021 and demonstrated strong generalization across all environmental conditions thanks to the expanded prompts explicitly requiring diverse environmental conditions. Notably, its performance in night conditions improved drastically (Dice = 0.7946), indicating that exposure to simulated low-light conditions during training enhanced robustness on real-world low-light images. Compared to the baseline real-world dataset model, this approach yielded superior segmentation performance across all test conditions, confirming the effectiveness of diffusion models in generating diverse, useful training data.

The most effective model combined the original 150 real-world images with the 1500 synthetic images from the expanded prompt, resulting in an overall Dice coefficient of 0.8612 which is the highest score among all models. This model consistently outperformed the others in all tested weather conditions, including night (Dice = 0.8385). The combination of real and synthetic data provided the model with both real-world grounding and the environmental diversity necessary for improved generalization. The ability to maintain high segmentation accuracy under varying conditions underscores the advantage of supplementing small, biased real-world datasets with synthetic images. Examples of accurate segmentation obtained using this model are shown in Fig. 5.

It is important to note that for the purpose of assisting visually impaired people, it is not necessary to detect every crosswalk pixel with absolute precision. Instead, ensuring that the majority of the crosswalk area is correctly identified is sufficient for safe navigation. While slightly lower recall values indicate that some crosswalk edges may not be fully covered in predictions, the high precision rates confirm a low rate of false positive detections, which is important for this application.

Although the 300 image high-quality test dataset was collected primarily for evaluation, ensuring a balanced representation of various weather and lighting conditions,

TABLE 3. Segmentation results.

Training dataset	Evaluation Subset	Size	Metrics Dice	IoU	Pix. Acc.	Precision	Recall
150 real-world images (initial dataset)	sunny	120	0.8045	0.7162	0.9390	0.8581	0.8106
	cloudy	60	0.8443	0.7541	0.9383	0.8460	0.8836
	rainy	60	0.8083	0.7082	0.9422	0.8390	0.8344
	night	60	0.1878	0.1316	0.8320	0.9499	0.1343
	all	300	0.6899	0.6053	0.9181	0.8702	0.6947
1500 synthetic images (obtained with base prompt)	sunny	120	0.6684	0.5709	0.8957	0.8781	0.6206
	cloudy	60	0.8004	0.7004	0.9313	0.9381	0.7366
	rainy	60	0.7730	0.6630	0.9271	0.8979	0.7211
	night	60	0.0670	0.0440	0.8112	0.9978	0.0443
	all	300	0.5954	0.5098	0.8922	0.9180	0.5486
1500 synthetic images (obtained with expanded prompt)	sunny	120	0.8017	0.6956	0.9261	0.9360	0.7350
	cloudy	60	0.8170	0.7131	0.9335	0.9393	0.7509
	rainy	60	0.7957	0.6876	0.9352	0.9117	0.7437
	night	60	0.7946	0.6857	0.9395	0.9762	0.6994
	all	300	**0.8021**	0.6955	0.9321	0.9398	0.7328
1500 synthetic images (expanded prompt) + 150 real-world images	sunny	120	0.8782	0.7942	0.9480	0.9031	0.8695
	cloudy	60	0.8880	0.8090	0.9507	0.9160	0.8696
	rainy	60	0.8547	0.7580	0.9456	0.8771	0.8546
	night	60	0.8385	0.7075	0.9421	0.9635	0.7291
	all	300	**0.8612**	0.7726	0.9469	0.9126	0.8385
300 real-world images from evaluation set	5-fold cross-validated		**0.9064**	0.8220	0.9632	0.9111	0.8947

TABLE 4. Generalization across different datasets.

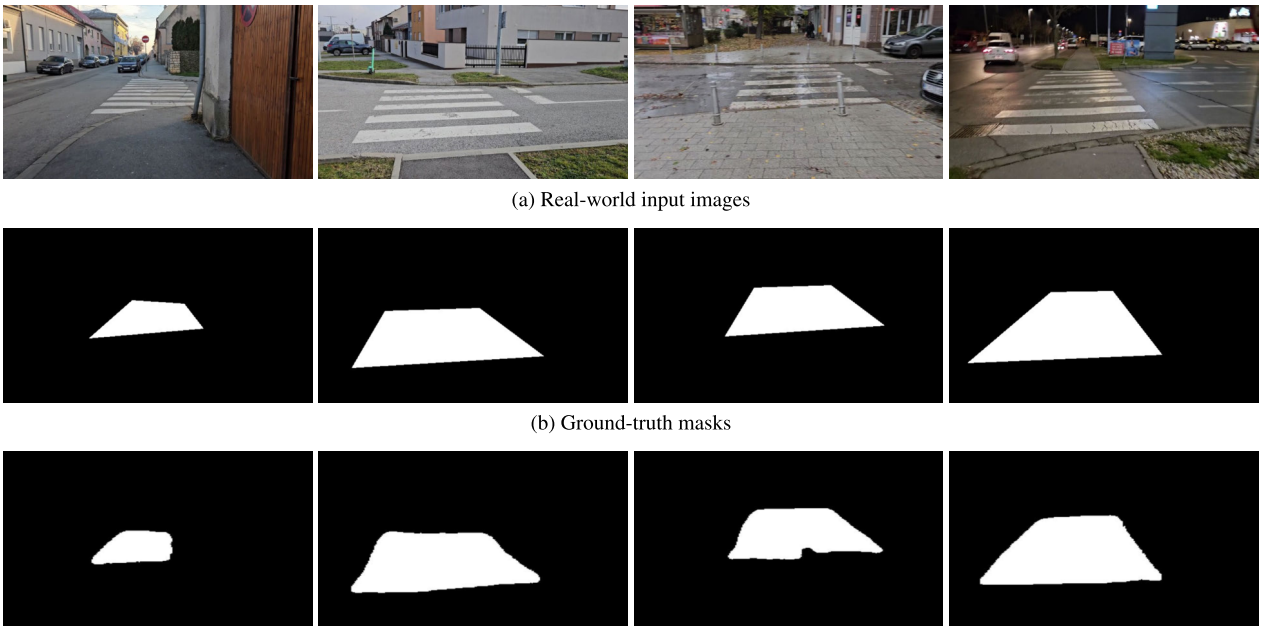
Model trained on	Test dataset	Dataset Size	Dice	IoU	Pix. Acc.	Precision	Recall
1500 synthetic images (expanded prompt) + 150 real-world images	D. Silva Medeiros dataset [26]	450	**0.8499**	0.7691	0.8683	0.9563	0.8043
	R. Cheng dataset (full) [27]	120	0.7393	0.6606	0.9631	0.8515	0.7280
	R. Cheng dataset (subset) [27]	105	**0.8303**	0.7453	0.9673	0.8794	0.8123
	Our dataset [23]	300	**0.8612**	0.7726	0.9469	0.9126	0.8385

we performed an additional training run solely on this dataset. The objective of this training run was to establish a strong performance benchmark using a high-quality dataset and 5-fold cross-validation. The results show that this model achieves a Dice score of 0.9064, indicating that a model trained on a highly diverse, real-world dataset can perform exceptionally well. However, in scenarios where real-world data is limited, training on a combination of real and diffusion-generated synthetic data provides a strong alternative, achieving comparable performance (Dice = 0.8612). This highlights the potential of Stable Diffusion-generated data in bridging the gap when large, diverse real-world datasets are impractical to obtain.

These findings confirm that diffusion-generated synthetic data can be effectively used to augment small real-world datasets, improving model generalization to challenging real-world scenarios. The proposed approach provides a scalable and adaptable solution for applications where data collection is limited or constrained by environmental conditions.

To further assess the geographic and contextual generalization of the proposed approach, we evaluated the model on two external datasets from different regions: the D. Silva Medeiros dataset (South America) [26] and the R. Cheng dataset (Europe) [27]. The Silva dataset consists of pedestrian-perspective imagery similar to our own, and the model achieved a Dice score of 0.8499, comparable to 0.8612 on our dataset. The Cheng dataset is also pedestrian-perspective; however, approximately 15 out of 120 images depict crosswalks at greater distances, resulting in more horizontally oriented patterns. Since our method is designed for detecting crosswalks at close range, typically a few meters ahead, the performance on these specific cases is somewhat lower. On the full Cheng dataset, the model achieved a Dice score of 0.7393, while on the subset with capturing approach and perspectives similar to our training data the score increased to 0.8303. Overall, these results (Table 4) indicate strong generalization to new geographic regions and unseen datasets under typical pedestrian perspectives, while highlighting that handling distant crosswalks falls slightly outside the current scope and represents a natural direction for future improvement.

Despite its strong overall performance, our method exhibits certain limitations that are important to acknowledge. First, the model was trained on pedestrian-perspective images in which crosswalks are located at a relatively short distance in front of the camera. Consequently, the model shows reduced accuracy on images where crosswalks appear farther away and are oriented horizontally with respect to the camera plane,



(a) Real-world input images

(b) Ground-truth masks

(c) Predictions

FIGURE 5. Successful predictions obtained using model trained on the mix of synthetic and real-world images.

as is often the case in vehicle-mounted viewpoints. Second, while the approach performs reliably in urban nighttime scenes with street lighting, its performance can degrade in extremely low-light environments without any artificial illumination, where crosswalk markings are barely visible. This limitation arises from the scarcity of such conditions in the training data. We note that these situations represent relatively rare edge cases, but acknowledging them helps define the current applicability boundaries of the proposed method and highlights directions for future improvement.

Finally, we acknowledge that the use of synthetic data introduces potential sources of bias and ethical considerations. Since the diffusion model was fine-tuned on a relatively small, geographically homogeneous dataset, the generated images tend to reflect similar urban characteristics, potentially limiting their diversity. Fine-tuning on broader and more diverse datasets could help mitigate this effect. Ethically, it is important to ensure that synthetic images are not misrepresented as real locations or individuals and that their artificial origin is transparent in any downstream use. While the current work focuses on demonstrating the feasibility of diffusion-based data augmentation, these considerations will be crucial for future large-scale deployments.

V. CONCLUSION

This paper investigates the feasibility of using Stable Diffusion-generated images to augment limited datasets for training deep learning models in specialized applications, such as crosswalk segmentation for assistive navigation systems for visually impaired individuals. Through fine-tuning, the Stable Diffusion model was able to generate highly realistic synthetic images that preserved the visual characteristics of real-world crosswalks while introducing diverse weather and lighting conditions. This approach provides an effective solution to the lack of first-person view (FPV) crosswalk datasets for deep learning applications.

Our findings demonstrate that dataset augmentation with synthetic images significantly enhances model performance. While training on only 150 real-world images resulted in a Dice coefficient of 0.6899, models trained with 1500 diffusion-generated images under diverse conditions achieved 0.8021, proving the importance of environmental diversity. Furthermore, combining real and synthetic data improved generalization, achieving a Dice score of 0.8612 and outperforming all other models across all tested conditions. These results confirm that augmenting small datasets with Stable Diffusion-generated images improves robustness, particularly in adverse weather and lighting scenarios.

The efficiency of Stable Diffusion also makes this approach practical for real-world deployment, as it enables dataset augmentation without requiring large-scale data collection or computationally expensive simulation pipelines. The proposed method offers a scalable solution for enhancing deep learning models in applications where annotated datasets are scarce, with potential for adaptation to other domains requiring synthetic data augmentation.

REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[2] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.

[4] C.-M. Feng, K. Yu, Y. Liu, S. Khan, and W. Zuo, "Diverse data augmentation with diffusions for effective test-time prompt tuning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 2704–2714.

[5] C. Xiao, S. X. Xu, and K. Zhang, "Multimodal data augmentation for image captioning using diffusion models," in *Proc. 1st Workshop Large Generative Models Meet Multimodal Appl.*, New York, NY, USA, Oct. 2023, pp. 23–33.

[6] B. Trabucco, K. G. Doherty, M. Gurinas, and R. Salakhutdinov, "Effective data augmentation with diffusion models," in *Proc. NeurIPS Workshop Synth. Data Gener. Generative AI*, 2023, pp. 1–22. [Online]. Available: https://openreview.net/forum?id=TTCIZunOVM

[7] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," 2021, *arXiv:2105.05233*.

[8] 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," in *Proc. Int. Conf. Mach. Learn.*, Dec. 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:245335086

[9] Z. Sordo, E. Chagnon, Z. Hu, J. J. Donatelli, P. Andeer, P. S. Nico, T. Northen, and D. Ushizima, "Synthetic scientific image generation with VAE, GAN, and diffusion model architectures," *J. Imag.*, vol. 11, no. 8, p. 252, Jul. 2025.

[10] V. Tümen and B. Ergen, "Intersections and crosswalk detection using deep learning and image processing techniques," *Phys. A, Stat. Mech. Appl.*, vol. 543, Apr. 2020, Art. no. 123510.

[11] Ö. Kaya, M. Y. Çodur, and E. Mustafaraj, "Automatic detection of pedestrian crosswalk with faster R-CNN and YOLOv7," *Buildings*, vol. 13, no. 4, p. 1070, Apr. 2023.

[12] G. Dogan and B. Ergen, "A new hybrid mobile CNN approach for crosswalk recognition in autonomous vehicles," *Multimedia Tools Appl.*, vol. 83, no. 26, pp. 67747–67762, Jan. 2024.

[13] J. Zhong, W. Feng, Q. Lei, S. Le, X. Wei, Y. Wang, and W. Wang, "Improved U-Net for zebra-crossing image segmentation," in *Proc. IEEE 6th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2020, pp. 388–393.

[14] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, Jul. 2022.

[15] S. Tian, M. Zheng, W. Zou, X. Li, and L. Zhang, "Dynamic crosswalk scene understanding for the visually impaired," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1478–1486, 2021.

[16] A. Kerim, F. Chamone, W. Ramos, L. S. Marcolino, E. R. Nascimento, and R. Jiang, "Semantic segmentation under adverse conditions: A weather and nighttime-aware synthetic data-based approach," in *Proc. 33rd Brit. Mach. Vis. Conf.*, London, U.K., 2022, pp. 1–14.

[17] G. J. M. Rosa, J. M. S. Afonso, P. D. Gaspar, V. N. G. J. Soares, and J. M. L. P. Caldeira, "Detecting wear and tear in pedestrian crossings using computer vision techniques: Approaches, challenges, and opportunities," *Information*, vol. 15, no. 3, p. 169, Mar. 2024.

[18] S. Mandia, A. Kumar, K. Verma, and J. K. Deegwal, "Vision-based assistive systems for visually impaired people: A review," in *Optical and Wireless Technologies*. Cham, Switzerland: Springer, 2023, pp. 163–172.

[19] A. Budrionis, D. Plikynas, P. Daniusis, and A. Indrulionis, "Smartphone-based computer vision travelling aids for blind and visually impaired individuals: A systematic review," *Assistive Technol.*, vol. 34, no. 2, pp. 178–194, Mar. 2022.

[20] E. J. Anthony and R. A. Kusnadi, "Computer vision for supporting visually impaired people: A systematic review," *Eng., Math. Comput. Sci. (EMACS) J.*, vol. 3, no. 2, pp. 65–71, May 2021.

[21] K. Romic, I. Galic, H. Leventic, and M. Habijan, "Pedestrian crosswalk detection using a column and row structure analysis in assistance systems for the visually impaired," *Acta Polytechnica Hungarica*, vol. 18, no. 7, pp. 25–45, 2021.

[22] K. Romic. (2025). *SD-Crosswalk-Augmentation Fine-Tuned Model*. [Online]. Available: https://huggingface.co/kromic/sd-crosswalk-augmentation

[23] K. Romic, H. Leventic, M. Habijan, and I. Galic, "Synthetic and real-world datasets for crosswalk segmentation under diverse weather and lighting conditions," *Data Brief*, vol. 61, Aug. 2025, Art. no. 111755.

[24] H. Zhu, H. Wei, B. Li, X. Yuan, and N. Kehtarnavaz, "A review of video object detection: Datasets, metrics and methods," *Appl. Sci.*, vol. 10, no. 21, p. 7834, Nov. 2020.

[25] D. Stursa, P. Rozsival, and P. Dolezel, "Efficient dataset extension using generative networks for assessing degree of coating degradation around scribe," *Frontiers Artif. Intell.*, vol. 7, Dec. 2024, Art. no. 1456844.

[26] E. T. Silva, F. Sampaio, L. C. da Silva, D. S. Medeiros, and G. P. Correia, "A method for embedding a computer vision application into a wearable device," *Microprocessors Microsyst.*, vol. 76, Jul. 2020, Art. no. 103086.

[27] R. Cheng. (2025). *Pedestrian Crosswalks Recognition (PCR) Public Database*. [Online]. Available: http://www.wangkaiwei.org/project.html

KREŠIMIR ROMIĆ was born in Osijek, Croatia. He received the M.Eng. degree in computer science and the Ph.D. degree from the Faculty of Electrical Engineering, Computer Science and Information Technology, J. J. Strossmayer University of Osijek, Osijek, Croatia, in 2011 and 2018, respectively. He is currently an Assistant Professor with the Faculty of Electrical Engineering, Computer Science and Information Technology. His research interests include computer vision and machine learning in image analysis and synthesis.

HRVOJE LEVENTIĆ was born in Osijek, Croatia. He received the M.Eng. degree in computer science from the Faculty of Electrical Engineering, Computer Science and Information Technology, J. J. Strossmayer University of Osijek, Osijek, Croatia, in 2012, and the joint Ph.D. degree from J. J. Strossmayer University of Osijek, and Ghent University, Belgium, in 2019. He is currently an Assistant Professor with the Faculty of Electrical Engineering, Computer Science and Information Technology. His research interests focus on biomedical image processing, machine learning, and large language models.

MARIJA HABIJAN received the M.Eng. degree in computer science from the Faculty of Electrical Engineering, Computer Science and Information Technology, J. J. Strossmayer University of Osijek, Osijek, Croatia, in 2015, and the joint Ph.D. degree from J. J. Strossmayer University of Osijek, and Ghent University, Belgium, in 2022. Since 2022, she has been a Postdoctoral Researcher with the Faculty of Electrical Engineering, Computer Science and Information Technology. Her research interests include machine learning, medical image segmentation, and computer vision, with a strong emphasis on developing advanced methods for the analysis of cardiac and vascular structures in medical imaging.

IRENA GALIĆ (Member, IEEE) was born in Osijek, Croatia. She received the Diploma degree in mathematics and computer science from J. J. Strossmayer University of Osijek, Croatia, in 1999, the M.Sc. degree in computer science from Saarland University, Saarbrücken, Germany, in 2004, and the Ph.D. degree from J. J. Strossmayer University of Osijek, in 2011. She is currently a Full Professor and the Head of the Department of Software Engineering with the Faculty of Electrical Engineering, Computer Science and Information Technology, J. J. Strossmayer University of Osijek. Her research interest includes visual computing and artificial intelligence.