IPTEK The Journal of Technology and Science, 35(3), 2024 (e/pISSN:2088-2033, 0853-4098)

DOI: 10.12962/j20882033.v35i3.22186 | Received 17 Dec, 2024; Revised 22 Feb, 2024; Accepted 27 Dec, 2024

ORIGINAL RESEARCH

SEMANTIC EDITING OF TRAFFIC NEAR-MISS AND ACCIDENT DATASET USING TUNE-A-VIDEO

Eka Alifia Kusnanti | Chastine Fatichah | Muhamad Hilmil Pradana*

Dept. of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Correspondence

*Muhamad Hilmil Pradana, Dept of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. Email: hilmi@its.ac.id

Present Address

Gedung Teknik Informatika, Jl. Teknik Kimia, Surabaya 60111, Indonesia

Abstract

Developing effective traffic monitoring systems for accident detection relies heavily on high-quality, diverse datasets. Many existing approaches focus on detecting anomalies in traffic videos. Still, they often fail to account for how varying environmental conditions, such as time of day, weather, or lighting, might influence the occurrence of near-misses or accidents. In this study, we explore the potential of Tune-A-Video to apply semantic editing techniques to an existing traffic near-miss and accident dataset. By modifying the visual environment, such as changing the time of day, weather, or lighting, we aim to generate realistic footage variations without altering the core events like near-miss incidents or accidents. This method enhances the dataset with more varied and realistic traffic conditions, improving its representativeness of real-world scenarios. The primary objective is not to create a new dataset but to assess the impact of semantic editing on the dataset's diversity and its effect on model performance. The results show that using Tune-A-Video for semantic editing can enrich the dataset, making it more suitable for training machine learning models. This approach helps improve the accuracy and robustness of computer vision models, particularly for traffic monitoring and accident detection applications, offering a promising tool for traffic safety systems.

KEYWORDS:

Accident, Near-Miss, Semantic Editing, Traffic Dataset, Tune-A-Video

1 | INTRODUCTION

The development of intelligent traffic monitoring systems has become increasingly critical as traffic-related accidents continue to be one of the leading causes of death and injury worldwide. These systems, particularly those focused on accident detection and near-miss identification, hold the potential to significantly improve road safety by providing real-time analysis of traffic incidents[1]. Near-misses, incidents where accidents are narrowly avoided, are important because they can show early signs of safety problems. Understanding and addressing these events can help prevent future accidents. However, these systems work best when high-quality datasets show real-world traffic situations accurately[2].

Existing traffic datasets often fall short in this regard. While many datasets provide sufficient coverage for accident detection, they frequently lack detailed annotations for near-miss scenarios, which are just as important for developing proactive safety systems[3]. Moreover, these datasets are usually limited in diversity, failing to capture the wide range of environmental and contextual factors that influence traffic incidents[4]. Conditions such as weather patterns, lighting variations, vehicle types, and road scenarios are rarely accounted for comprehensively. This lack of diversity reduces the robustness of machine learning models, making them less effective when applied to real-world traffic environments[5].

Researchers have turned to data augmentation techniques to address these challenges to enhance dataset diversity and size. Traditional augmentation methods, such as rotation, scaling, or flipping, are widely used in image-based tasks to prevent overfitting and improve generalization[6]. However, these techniques are less effective for video data, where maintaining the semantic context of events such as accidents or near-misses is crucial. Basic transformations may distort or obscure critical details, limiting their usefulness for traffic monitoring and safety applications[7].

Semantic editing has emerged as a promising alternative for augmenting video datasets[8]. Unlike traditional methods, semantic editing allows for precise, controlled changes to specific aspects of a video, such as altering weather conditions, adjusting lighting, or changing the time of day. Importantly, this approach ensures that the core events within the footage remain intact, preserving their relevance for training machine learning models[9]. By enabling the generation of realistic variations, semantic editing creates datasets that better reflect the complexities of real-world traffic scenarios, thereby improving the robustness and generalizability of computer vision models[10].

Among the various tools available for semantic editing, Tune-A-Video stands out as a highly effective method for video augmentation[11]. Tune-A-Video leverages generative models to modify visual attributes of traffic footage while maintaining the integrity of the key events, such as near-misses or accidents[12]. Using techniques like conditional style translation[13], Tune-A-Video can simulate diverse traffic environments, including variations in weather, time of day, and road conditions. These realistic modifications introduce new dimensions of variability to existing datasets, enhancing their utility for training machine learning models[14].

Although it has potential, using Tune-A-Video for semantic editing in traffic datasets has not been widely explored. Most existing studies focus on analyzing traffic incidents using static datasets, with minimal emphasis on augmenting data to reflect real-world variability[15]. When augmentation is used, it often relies on basic image-processing techniques that do not adequately capture the complexity of traffic events. This gap in the literature highlights the need for advanced methods to generate realistic, diverse video samples to improve the performance of traffic monitoring and safety systems[16].

This study aims to address these gaps by exploring using Tune-A-Video for semantic editing to augment a traffic near-miss and accident dataset. The primary objective is to enhance dataset diversity by modifying semantic elements like weather conditions, lighting, and time of day while preserving the core events of interest. By generating realistic dataset variations, this research seeks to create a more comprehensive representation of real-world traffic conditions, improving the robustness of machine learning models for accident detection and traffic safety applications. Additionally, this study demonstrates how Tune-A-Video can be a practical and effective tool for dataset augmentation, especially in domains where collecting diverse real-world data is challenging and resource-intensive. Through this approach, we aim to contribute to the development of more effective traffic monitoring systems capable of operating reliably across a wide range of conditions. .

2 | PREVIOUS RESEARCHES

Recent advancements in traffic accident and near-miss detection increasingly rely on video datasets, where annotation quality and dataset diversity significantly impact model performance[17]. Traditional data augmentation techniques, such as rotation, flipping, and scaling, are widely used for image datasets to improve model generalization[18]. However, these methods are less effective for video data due to the complexity of temporal information and contextual dependencies essential for accurately detecting incidents like accidents and near-misses[3]. To address this gap, advanced augmentation approaches that consider the temporal dynamics of videos, such as dynamic object manipulations, have been explored[19]. Semantic editing has emerged as a promising method for augmenting video datasets. This technique modifies specific visual elements in a video, such as lighting, weather, and vehicle types, without altering the core events, such as accidents or near-misses. By simulating diverse[20] real-world conditions, semantic editing can significantly improve dataset variety and enhance the robustness of machine learning models trained on these datasets[21].

One significant development in this field is Conditional Style Translation (CST), a generative method that transfers specific styles between video frames[13]. CST has introduced variations in visual attributes, such as lighting and environmental conditions while maintaining the original sequence of events. For instance, previous studies applying CST to datasets like DADA-2000 focused on traffic accidents and near-miss events demonstrated significant improvements in model accuracy. By increasing dataset diversity, CST enhanced the generalization capabilities of models for traffic risk classification, with positive accuracy margins observed during cross-validation[17]. However, CST is not without limitations. Lighting adjustments, for example, can sometimes produce inconsistent results, such as frames that are overly bright or too dark, reducing the realism of augmented data. These challenges underscore the need for more refined augmentation methods that ensure consistent and realistic modifications[13].

Building on these advancements, Tune-A-Video[12] represents a recent breakthrough in video augmentation. This method enables controlled manipulation of video attributes, such as vehicle types, road conditions, and background elements, while preserving the integrity of key events. By generating realistic variations of traffic scenarios, Tune-A-Video offers a more effective approach for enhancing dataset diversity and improving the robustness of traffic monitoring systems.

Despite the potential of generative models and semantic editing techniques, most prior research has focused primarily on accident detection, often neglecting near-miss scenarios or relying on basic augmentation methods that fail to capture the complexity of real-world traffic conditions[22]. Furthermore, the application of Tune-A-Video for augmenting traffic datasets remains underexplored, particularly for scenarios involving accidents and near-misses.

This study aims to address these gaps by applying Tune-A-Video for semantic editing to augment a traffic near-miss and accident dataset. By introducing realistic variations through controlled modifications of semantic elements such as lighting, weather, and vehicle types, this research seeks to enhance dataset diversity, improve model generalizability, and support the development of more robust systems for accident detection and traffic safety applications.

3 | METHOD

This section details the methods employed in this study, emphasizing using Tune-A-Video for semantic editing to enhance the traffic near-miss and accident dataset.

3.1 | Tune-A-Video

Tune-A-Video is an advanced text-to-video model that generates video content based on textual descriptions[12]. The model uses a generative approach to produce video sequences that align with the details specified in the text, including environmental factors, vehicle appearances, and traffic conditions[23]. The flexibility of Tune-A-Video allows for significant control over the generated content, enabling the creation of diverse traffic scenarios, which are crucial for training models to detect accidents and near-misses under a variety of real-world conditions[3].

For example, a text input such as “a highway scene at night with heavy rain” would prompt Tune-A-Video to generate a video that accurately reflects these specific conditions, modifying weather, vehicle behavior, and road conditions. This ability to generate realistic traffic situations based on text descriptions helps expand the diversity of the dataset without the need to capture new footage. As a result, Tune-A-Video is particularly useful for augmenting datasets to cover a broader range of traffic scenarios that may not be present in the original data[12].

The architecture of Tune-A-Video typically involves a deep learning-based generative model that incorporates text-conditioned video generation. The model takes a textual description as input and generates corresponding video frames that match the description’s context. The process relies on advanced unsupervised image-to-image translation techniques to ensure the generated video is coherent and realistic. Figure 1 illustrates the architecture of Tune-A-Video, showing the flow from text input to the generated video output, highlighting key components such as the text encoder, video generator, and discriminator. These components work together to create videos visually consistent with the input text description[12].

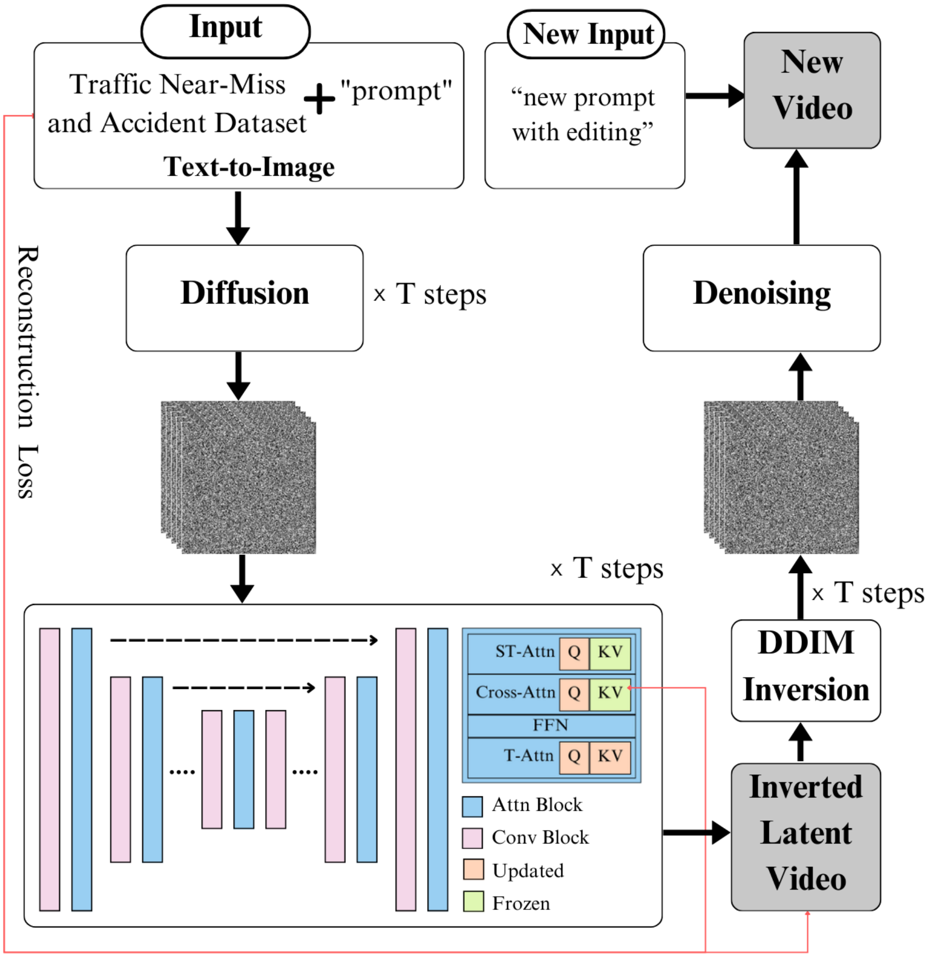

FIGURE 1 The architecture of Tune-A-Video.

3.2 | Semantic editing

Semantic editing involves the targeted manipulation of specific elements within a video while preserving the core events and objects[24]. This study uses semantic editing to modify aspects such as lighting, weather, and time of day without altering the fundamental traffic event, such as a near-miss or accident. This approach allows for the simulation of different environmental conditions, such as changing from day to night or modifying the weather, which is important for creating a diverse and robust dataset[25].

A key application of semantic editing is altering the lighting and environmental factors, for example, transitioning a scene from daylight to nighttime or adding weather effects like rain or fog[26]. By adjusting these elements, the augmented video can reflect various conditions under which traffic accidents or near-misses might occur. Simulating these variations helps improve the model's performance by exposing it to a wider range of driving conditions[20].

In this study, semantic editing is applied using Tune-A-Video, where text descriptions are used to modify the visual elements of the video. Figure 2 illustrates an overview of this approach. These edits ensure the creation of diverse training data that can help the model generalize to different traffic scenarios, improving its ability to detect accidents in varied conditions.

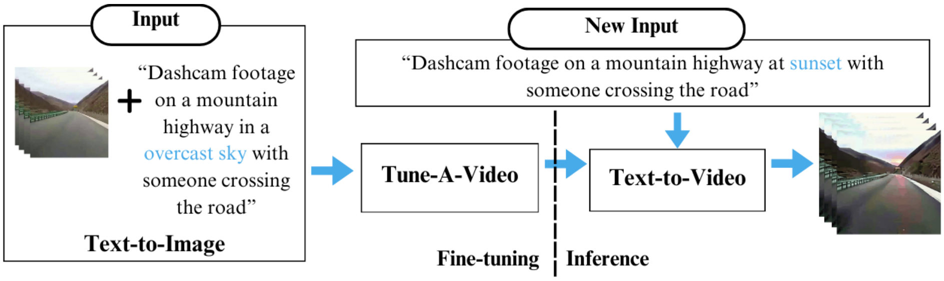
Input
"Dashcam footage on a mountain highway in a overcast sky with someone crossing the road"
Text-to-Image
New Input
"Dashcam footage on a mountain highway at sunset with someone crossing the road"
Tune-A-Video
Text-to-Video
Fine-tuning
Inference

FIGURE 2 An overview of semantic editing using Tune-A-Video.

4 | RESULT AND DISCUSSION

This section presents the results of applying Tune-A-Video for semantic editing on the traffic near-miss and accident dataset, followed by a discussion of the impact and implications of these results.

4.1 | Dataset and Implementation Settings

In this study, we utilized the DADA-2000[27] dataset along with additional near-miss data to explore the effectiveness of Tune-A-Video (TAV) for semantic editing. The primary goal was to generate realistic environmental variations, such as changes in lighting, time of day, and weather, while ensuring that the core events, like near-misses and accidents, remained unchanged. To enhance the dataset, we extended the start and end times of accident durations to fully capture all ego-motions during the incidents. The dataset was processed using Tune-A-Video, which enabled precise and controlled modifications to visual attributes. This implementation was performed to create realistic augmentations, enhancing the diversity of the dataset without distorting key events. The outputs were evaluated based on two metrics: Frame Consistency: Measures smoothness between consecutive frames, which ensures temporal coherence. Textual Alignment: Assess how accurately the generated videos reflect the intended semantic changes (e.g., lighting or weather modifications).

4.2 | Results of Semantic Editing Using Tune-A-Video

Tune-A-Video generates high-quality and realistic edits of traffic scenes, demonstrating its ability to enhance datasets while preserving important traffic events. As shown in Figure 3 , it effectively applied various environmental changes, such as transitioning from daylight to sunset and from clear skies to cloudy conditions. Despite these adjustments, key traffic events, such as near-misses or accidents, remained visually clear and easily interpretable. This consistency is essential for training accurate machine learning models for accident detection and traffic safety analysis.

A key feature of the modifications generated by Tune-A-Video is the consistent lighting and smooth transitions between frames. For example, in Figure 4 , the lighting naturally changes across the frames, moving seamlessly from bright daylight to the warm tones of sunset. Similarly, as shown in Figure 4, when viewed frame by frame, the lighting remains steady, and the transitions between frames are smooth, with no sudden changes that could disrupt the flow of the video. These smooth transitions ensure no abrupt changes, which might otherwise distort the dataset. Such consistency in lighting is important for creating a realistic dataset that reflects the dynamic and varied environmental conditions seen in real-world traffic situations. When lighting shifts from bright to overcast or daytime to nighttime, the edits appear natural and believable, further enhancing the realism of the video data.

Measurable results also support the performance of Tune-A-Video. The Frame Consistency score of 0.9695 highlights the high level of temporal coherence in the augmented videos. This means that the modified frames transition smoothly without noticeable jumps or inconsistencies, maintaining the natural flow of the video. Such smooth transitions are essential for keeping the augmented footage realistic, especially for training machine learning models that need high-quality and consistent data. Additionally, the Textual Alignment score of 0.3310 shows that the changes made to the video, such as lighting or weather adjustments, were applied accurately as intended. This result confirms that the edits were precise and did not compromise the original content.

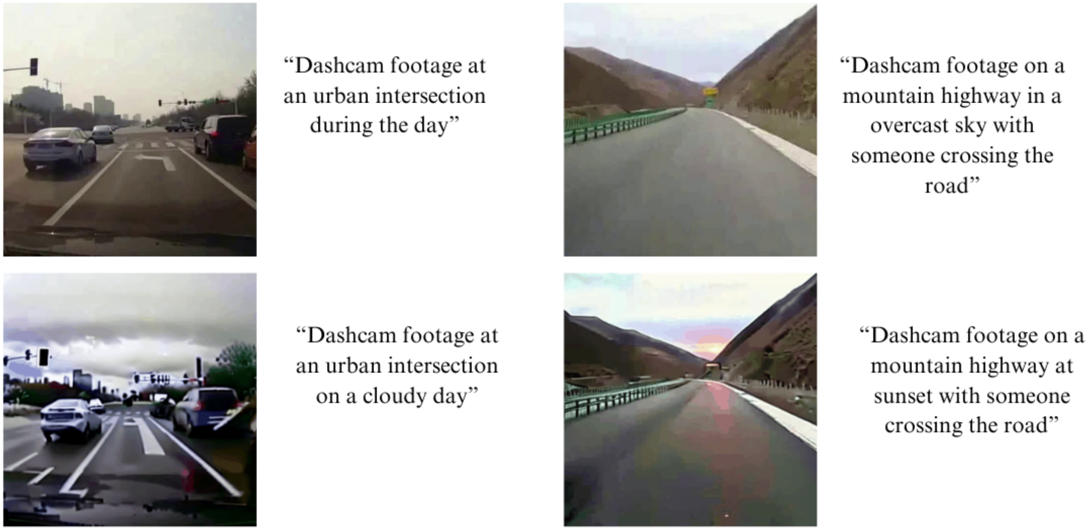
"Dashcam footage at an urban intersection during the day"
"Dashcam footage on a mountain highway in a overcast sky with someone crossing the road"
"Dashcam footage at an urban intersection on a cloudy day"
"Dashcam footage on a mountain highway at sunset with someone crossing the road"

FIGURE 3 Sample results of Tune-A-Video: Semantic editing of traffic near-miss and accident.

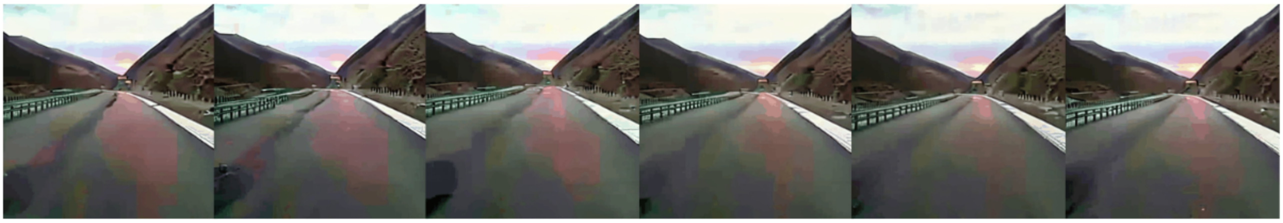

FIGURE 4 Frame-by-frame visualization of lighting transitions in Tune-A-Video outputs.

By generating realistic and coherent video outputs with consistent lighting and smooth transitions, Tune-A-Video ensures the key traffic events remain intact and easily understood. The measurable results further confirm the quality of the modifications, showing that Tune-A-Video can create diverse and realistic variations of traffic footage. This makes it a valuable tool for improving the robustness and variety of traffic datasets used in machine learning and computer vision research.

4.3 | Comparison with Conditional Style Translation (CST)

To assess the performance of Tune-A-Video (TAV) against other methods, Conditional Style Translation (CST) was applied to the same dataset for comparison. As shown in Figure 5 , CST can successfully introduce changes to the visual environment, such as adjustments in lighting or weather. However, CST's notable limitation is its inability to maintain consistent lighting and uniform appearance across frames. In some outputs, CST produces areas with inconsistent brightness, where certain regions appear overly bright while others remain unnaturally dark. This inconsistency can reduce the overall realism of the modified videos.

The quantitative evaluation further highlights these differences. CST achieves a Frame Consistency score of 0.9834, slightly higher than TAV's score of 0.9520. This indicates that CST performs marginally better in maintaining temporal smoothness across video frames. However, when examining Textual Alignment, TAV outperforms CST with a score of 0.3321, compared to CST's 0.2942. This suggests that TAV is more accurate in applying the intended semantic changes, such as lighting or weather modifications while ensuring the visual coherence of key events remains intact.

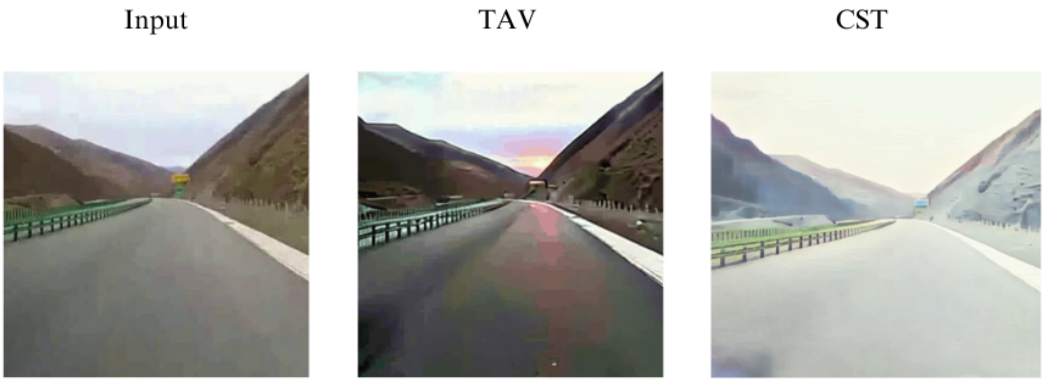

FIGURE 5 Comparison of Tune-A-Video with CST.

These findings indicate that while CST offers smoother transitions between frames, it struggles with achieving precise semantic edits and consistent lighting adjustments. In contrast, TAV produces outputs that more effectively align with the desired modifications, preserving the integrity of critical events, as demonstrated in Figure 5 . This makes TAV a more reliable tool for enhancing datasets where visual clarity and semantic consistency are crucial, such as traffic near-miss and accident scenarios.

4.4 | Analysis of Results

The findings demonstrate that Tune-A-Video offers a more effective semantic editing approach than CST. TAV produces realistic and uniform environmental changes while maintaining the integrity of core traffic events. The outputs look natural and ensure that critical visual details, such as vehicles, road conditions, and near-miss incidents, remain unaltered.

The slightly lower Frame Consistency score for Tune-A-Video compared to CST can be attributed to its focus on accurately applying semantic modifications, which sometimes introduces minor variations across frames. However, the higher Textual Alignment score clearly highlights its advantage in producing outputs that match the intended edits. By generating realistic variations in lighting, weather, and time of day, Tune-A-Video enhances the diversity of traffic datasets. This improvement is particularly valuable for training machine learning models, enabling better generalization across varying environmental conditions.

5 | CONCLUSION

In this study, we utilized Tune-A-Video for semantic editing of traffic near-miss and accident datasets, focusing on modifying environmental factors such as lighting and weather while preserving the integrity of critical events. The results show that Tune-A-Video produces realistic and consistent outputs, with smooth transitions and uniform lighting across frames, while keeping the core events visually clear and easy to understand. Quantitative metrics further validate its performance, with a Frame Consistency score of 0.9695, indicating smooth temporal coherence, and a Textual Alignment score of 0.3310, confirming an accurate representation of intended edits. These findings highlight the effectiveness of Tune-A-Video in producing diverse, high-quality augmentations that address the limitations of traditional methods, making it a valuable tool for enhancing dataset variability. By enabling realistic modifications that reflect real-world conditions, Tune-A-Video improves dataset representativeness, supporting the development of robust traffic monitoring and accident detection systems.

ACKNOWLEDGMENT

We would like to express our gratitude to the providers of the DADA-2000 dataset and the additional near-miss data, whose contributions were essential to this study. We also extend our heartfelt thanks to everyone and every institution involved in supporting this research, providing guidance, resources, and encouragement that enabled us to complete this work successfully.

CREDIT

Eka Alifia Kusnanti: Methodology, Software, Writing. **Chastine Fatichah:** Supervision, Validation. **Muhamad Hilmil Muchtar Aditya Pradana:** Conceptualization, Supervision.

References

1. Apostolovski N, Trajanovski N, Chavdar M, Kartalov T, Gerazov B, Ivanovski Z. Deep Learning Based Multimodal Information Fusion for Near-Miss Event Detection in Intelligent Traffic Monitoring Systems. In: Complex Systems: Spanning Control and Computational Cybernetics: Applications: Dedicated to Professor Georgi M. Dimirovski on his Anniversary Springer; 2022.p. 357–388.

2. Niu Y, Fan Y, Ju X. Critical review on data-driven approaches for learning from accidents: comparative analysis and future research. Safety science 2024;171:106381.

3. Yang G, Sarkar A, Ridgeway C, Thapa S, Jain S, Miller A. Using Artificial Intelligence/Machine Learning Tools to Analyze Safety, Road Scene, Near-Misses and Crashes. National Surface Transportation Safety Center for Excellence; 2024.

4. Sohail A, Cheema MA, Ali ME, Toosi AN, Rakha HA. Data-driven approaches for road safety: A comprehensive systematic literature review. Safety science 2023;158:105949.

5. Azfar T, Li J, Yu H, Cheu RL, Lv Y, Ke R. Deep learning-based computer vision methods for complex traffic environments perception: A review. Data Science for Transportation 2024;6(1):1–27.

6. Alomar K, Aysel HI, Cai X. Data augmentation in classification and segmentation: A survey and new strategies. Journal of Imaging 2023;9(2):46.

7. Abdel-Aty M, Wang Z, Zheng O, Abdelraouf A. Advances and applications of computer vision techniques in vehicle trajectory generation and surrogate traffic safety indicators. Accident Analysis & Prevention 2023;191:107191.

8. Patel AS, Vyas R, Vyas O, Ojha M. A study on video semantics; overview, challenges, and applications. Multimedia Tools and Applications 2022;81(5):6849–6897.

9. Gao Z, Chen X, Xu J, Yu R, Zhang H, Yang J. Semantically-Enhanced Feature Extraction with CLIP and Transformer Networks for Driver Fatigue Detection. Sensors 2024;24(24):7948.

10. Muhammad K, Hussain T, Ullah H, Del Ser J, Rezaei M, Kumar N, et al. Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. IEEE Transactions on Intelligent Transportation Systems 2022;23(12):22694–22715.

11. Gu J, Fang Y, Skorokhodov I, Wonka P, Du X, Tulyakov S, et al. VIA: A Spatiotemporal Video Adaptation Framework for Global and Local Video Editing. arXiv preprint arXiv:240612831 2024;.

12. Wu JZ, Ge Y, Wang X, Lei SW, Gu Y, Shi Y, et al. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023. p. 7623–7633.

13. Pradana H, Dao MS, Zettsu K. Augmenting ego-vehicle for traffic near-miss and accident classification dataset using manipulating conditional style translation. In: 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA) Sydney, Australia: IEEE; 2022. p. 1–8.

14. Uhlig S, Alkhasli I, Schubert F, Tschöpe C, Wolff M. A review of synthetic and augmented training data for machine learning in ultrasonic non-destructive evaluation. Ultrasonics 2023 September;134:107041.

15. Chen D, Zhu M, Yang H, Wang X, Wang Y. Data-driven Traffic Simulation: A Comprehensive Review. IEEE Transactions on Intelligent Vehicles 2024;9(4):4730–4748. https://ieeexplore.ieee.org/document/10440492.

16. Razi A, Chen X, Li H, Wang H, Russo B, Chen Y, et al. Deep learning serves traffic safety analysis: A forward-looking review. IET Intelligent Transport Systems 2023;17(1):22–71.

17. Rocky A, Wu QJ, Zhang W. Review of Accident Detection Methods Using Dashcam Videos for Autonomous Driving Vehicles. IEEE Transactions on Intelligent Transportation Systems 2024;25(8):8356–8374.

18. Garcea F, Serra A, Lamberti F, Morra L. Data augmentation for medical imaging: A systematic literature review. Computers in Biology and Medicine 2023;152:106391.

19. Eze C, Crick C. Learning by Watching: A Review of Video-based Learning Approaches for Robot Manipulation. arXiv preprint arXiv:240207127 2024;p. 1–26.

20. Rabbi ABK, Jeelani I. AI integration in construction safety: Current state, challenges, and future opportunities in text, vision, and audio based applications. Automation in Construction 2024;164:105443.

21. Wu R, Yang T, Sun L, Zhang Z, Li S, Zhang L. Seesr: Towards semantics-aware real-world image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2024. p. 25456–25467.

22. Mustapha A, Abdul-Rani AM, Saad N, Mustapha M. Advancements in Traffic Simulation for Enhanced Road Safety: A Review. Simulation Modelling Practice and Theory 2024;p. 103017.

23. Sun W, Tu RC, Liao J, Tao D. Diffusion model-based video editing: A survey. arXiv preprint arXiv:240707111 2024;p. 1–23.

24. Zhao L, Zhang Z, Nie X, Liu L, Liu S. Cross-Attention and Seamless Replacement of Latent Prompts for High-Definition Image-Driven Video Editing. Electronics 2023;13(1):1–14.

25. Testolina P, Barbato F, Michieli U, Giordani M, Zanuttigh P, Zorzi M. Selma: Semantic large-scale multimodal acquisitions in variable weather, daytime and viewpoints. IEEE Transactions on Intelligent Transportation Systems 2023;24(7):7012–7024.

26. Suryanto N, Adiputra AA, Kadiptya AY, Le TTH, Pratama D, Kim Y, et al. Cityscape-Adverse: Benchmarking Robustness of Semantic Segmentation with Realistic Scene Modifications via Diffusion-Based Image Editing. arXiv preprint arXiv:241100425 2024;p. 1–19.

27. Fang J, Yan D, Qiao J, Xue J, Yu H. DADA: Driver attention prediction in driving accident scenarios. IEEE transactions on intelligent transportation systems 2021;23(6):4959–4971.

How to cite this article: Kusnanti E.A., Fatichah C., Pradana M.H., (2024), Semantic Editing of Traffic Near-Miss and Accident Dataset Using Tune-A-Video, *IPTEK The Journal of Technology and Science*, *35(3):177–185*.