Evaluating Image-to-Image Translation Techniques for Simulating Physical Conditions of Traffic Signs

Rung-Ching Chen, Ming-Zhong Fan, William Eric Manongga, and Chayanon Sub-r-pa *

Department of Information Management, Chaoyang University of Technology, Taichung, Taiwan
Email: crching@cyut.edu.tw (R.-C.C.); s11214605@gm.cyut.edu.tw (M.-Z.F.); s11014907@cyut.edu.tw (W.E.M.); t5220317@gm.cyut.edu.tw (C.S.)
*Corresponding author

***Abstract*—Traffic signs are vital in providing important information to drivers, ensuring their safety, and helping them follow the road rules. Object detection algorithms like You Only Look Once (YOLO) are used in autonomous vehicles to monitor traffic sign information. However, most object detection research focuses on identifying traffic signs rather than their physical condition. One major issue with the existing dataset is the lack of data on damaged traffic signs for training, which could adversely affect the performance of the object detection algorithm. To address this problem, our paper comprehensively reviews the Image-to-Image (I2I) algorithm to modify existing traffic sign images to showcase different physical statuses (normal and damaged). We conduct experiments using state-of-the-art unpaired image-to-image translation techniques, UNet Vision Transformer cycle-consistent Generative Adversarial Network (UVCGAN) v2, and Energy-Guided Stochastic Differential Equations (EGSDE) to translate normal and damaged traffic sign images. Our experimental results are evaluated using Fréchet Inception Distance (FID) and side-by-side image comparison. We analyze and discuss possible and future improvements.**

***Keywords*—traffic sign detection, image generative, Image-to-Image (I2I), Generative Adversarial Networks (GANs), Cycle Generative Adversarial Network (CycleGAN), diffusion model**

I. Introduction

Traffic signals and signs play a crucial role in maintaining road safety. They provide necessary guidance and information to drivers and pedestrians, helping regulate traffic flow and prevent accidents. The Detection and Recognition of Traffic Signs (TSDR) [1] and interpretation of such signs are essential to the decision-making processes of all drivers and autonomous vehicles.

It is essential to continuously monitor the status of roads and traffic signs to ensure safe driving. However, monitoring extensive road networks can be challenging. In this regard, computer vision technology offers a promising alternative for consistent monitoring. Integrating computer vision technology into traffic sign monitoring systems can provide accurate and continuous monitoring.

However, damaged, faded, obscured, or vandalized traffic signs can usually be seen in the road network, affecting drivers and autonomous vehicles [2]. Poor visibility and legibility due to this factor can significantly increase road risks, which is a practical issue [3].

Traffic signs can lead to illegibility, fading, or damage, making it difficult for drivers to read and respond to them accurately. For the safety of road users, it is crucial to monitor and maintain traffic signs regularly. However, the traditional approach [4] of inspecting each sign is labor-intensive, time-consuming, and costly.

The effectiveness of computer vision systems heavily depends on the quantity and variety of data used to train them. However, there is currently a lack of datasets that include damaged traffic signs. This shortage of data limits the ability of these systems to accurately identify and classify damaged signs, which poses a significant challenge in developing reliable traffic sign damage monitoring systems.

Generative models like Generative Adversarial Network (GAN) [5], Deep Convolutional Generative Adversarial Network (DGAN) [6], Wasserstein Generative Adversarial Network (WGAN) [7], and others have played a key role in overcoming the problem of insufficient data by creating synthetic images that can be used to train models. However, these models are limited because they are designed to generate images of specific objects, while object detection training requires images of entire scenes with multiple objects.

The Diffusion Model (DM) [8] is an advanced generative approach that gradually creates high-quality images using conditional or text prompts. Although text-to-image [9] can generate scenes of road traffic, controlling the design image can be difficult, and many images have problems with small details. Using images from DM to train object detection can lead the model to learn unrealistic patterns that can affect accuracy.

In object detection applications [10–12], image generation can increase the data for the training by

Manuscript received February 15, 2024; revised February 29, 2024; accepted April 26, 2024; published September 5, 2024.

generating or modifying only a part of an existing image. For example, to create a traffic image with a damaged traffic sign, crop it from the labeled boundary box, generate Image-to-Image (I2I), and then replace it in the original image. The process of modifying traffic sign conditions in road traffic images is illustrated in Fig. 1.

This study used advanced I2I translation techniques to modify images of traffic signs from normal to damaged states, following the Fig. 1 approach. The research comprises a detailed experimental analysis of the state-of-the-art unpaired I2I translation techniques, UNet Vision Transformer cycle-consistent Generative Adversarial Network (UVCGAN) v2 [13] and Energy-Guided Stochastic Differential Equations (EGSDE) [14].

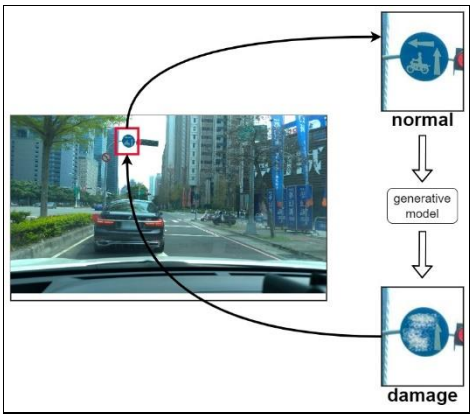

Fig. 1, Modify traffic sign from the road traffic image.

The contributions of this research include (1) Comprehensive experiment and analysis of the result when using UVCGANv2 and EGSDE to translate between normal and damaged traffic sign images. (2) Analyze and discuss the issue in the experiment for future improvement.

The paper is structured as follows: Section II presents the related work, while Section III outlines the research methodology, including details of the dataset and evaluation metric. Section IV provides the experimental results in Fréchet Inception Distance (FID) and image comparison, and Section V discusses the issues observed during the experiment and results for future improvement. Finally, Section VI concludes the research.

II. Related Work

A large dataset is required for model training to improve the accuracy of traffic signals and signal detection [15]. However, the dataset may be required to handle specific scenarios, such as damaged traffic signs. More data for such cases can result in adequate model training, making it easier for models to handle real-world situations. Therefore, adding more data that includes information on damaged traffic signs is important to improve model training and enhance their performance in real-world scenarios.

Data augmentation [16] methods have been employed to counteract data insufficient issues. These techniques artificially expand the dataset, enhancing the diversity of training samples. Traditional methods include geometric transformations, color space adjustments, and random cropping, which introduce variations in the dataset that mimic real-world conditions to an extent.

Recent advancements in image generation techniques have created new opportunities for data augmentation. Generative models are one of the most successful methods, particularly the GAN-based approach [5–7]. These models have been able to produce realistic images by learning to generate new ones that are nearly identical to real images. Image generative model has provided a way to create diverse training data artificially.

For example, Zhao *et al.* [17] uses Category-consistent and Relativistic Diverse Conditional Generative Adversarial Network (CRDCGAN) to increase the data for the training, improving the accuracy of classifying small-scale rock images, and Sandfort *et al.* [18] uses Cycle Generative Adversarial Network (CycleGAN) to enhance CT image segmentation performance. This approach addresses the gap in traditional data augmentation techniques, offering a more nuanced and comprehensive method for dataset expansion.

Generating datasets of scene images, such as road traffic images, is a challenging task for improving object detection training because the generated image may require more control and detailed realism. One potential solution is image modification or I2I [19] for specific objects since most images are real-world.

I2I translation [13, 14, 19] has great potential for use in Data Augmentation [16]. The I2I approach can convert images from one domain to another, such as transforming a clear, undamaged traffic sign into one that appears aged or damaged. This ability is crucial for training models to recognize and interpret traffic signs that are faded, vandalized, or deteriorated due to environmental factors.

CycleGAN [20] is a significant technique in unpaired image-to-image translation. It can convert images from one domain to another without corresponding image pairs. This innovative approach employs two generators and discriminators along with a cycle consistency loss to ensure that the crucial features of the original domain are preserved during translation. CycleGAN framework has opened up new possibilities in various fields, such as artistic style transfer and transforming real-world scenarios. It enables realistic and contextually appropriate image transformations.

UVCGANv2 [13] is a new image translation approach based on CycleGAN. It improves the generator's network architecture by combining UNet [21] and Vision Transformer (ViT) [22]. This hybrid approach enhances the model's ability to accurately translate images with more detailed and contextual information, which is particularly important for complex tasks such as traffic sign transformation.

UVCGANv2 has architectural improvements and refined training techniques compared to its predecessor, UVCGAN [23]. These updates allow UVCGANv2 to outperform other generative models in generating realistic image translations, making it highly suitable for applications where detail and accuracy are crucial.

The I2I using DM-based EGSDE [14] is a method for translating images from one domain to another without needing paired images. It achieves this by using a pre-trained energy function that enhances the realism and faithfulness of the translation process. This energy function guides a pre-trained Stochastic Differential Equation (SDE) to infer the most accurate translation between the source and target domains.

However, UVCGANv2 [13] and EGSDE [14] experimented with male-to-female, cat-to-dog, selfie-to-anime, and remove-glasses tasks, which differ from normal-to-damaged traffic sign tasks. To evaluate the performance of UVCGANv2 and EGSDE in translating traffic signs from their normal state to a damaged state, we conducted an experiment using the official implementation code provided in the original papers.

III. Research Methodology

Our research objective is to find a translation model that can effectively create normal and damaged traffic signs. To achieve this, we are experimenting with two state-of-the-art models: UVCGANv2 and EGSDE. It is important to note that UVCGANv2 is specifically designed for 256×256-pixel images, while the official implementation of EGSDE provides code that customizes each image size for each dataset. Therefore, we have adjusted the settings of EGSDE to produce images of the same size for a fair comparison.

A. UVCGANv2 Training Detail

Our experiment uses UVCGANv2 source code and training parameters from the official website. The training involves a two-step. In the first step, we pre-trained the generator self-supervised for image inpainting [24]. In contrast, the second step is the actual training of the unpaired I2I translation networks, starting from the pre-trained generators.

The generators are pre-trained on image inpainting tasks. This task is similar to the Bidirectional Encoder Representations from Transformers (BERT) pretraining [25]. For the inpainting task, input images of size 256×256 pixels are divided into a grid of patches at 32×32 pixels. Each patch is masked with a probability of 40%. The masking is performed by zeroing out pixel values. The generator is responsible for recovering the original unmasked image.

In the second step, translation training using the Adam optimizer with a beta equal to (0.5, 0.99), a learning rate of 1×10^{-4}, and training for 500 epochs. We have used the default values for hyperparameters from the male-to-female tasks. Since our tasks differ from those in UVCGANv2's research, we might need to explore the hyperparameters in the future.

B. EGSDE Training Detail

We follow the official implementation for EGSDE. The training process involves two steps: the diffusion model training and the SDE training. It's important to note that the diffusion model training process differs from the general deep learning approach. Each iteration of the diffusion model includes "*n*" images with a random "*t*" value for each image. The training process is repeated until the design iterator or target loss is reached.

Our experiment used the Ablated Diffusion Model (ADM) [26] instead of Denoising Diffusion Probabilistic Models (DDPM) for EGSDE, showing better results in the original paper. However, the training and usage of ADM and DDPM [27] are similar. In this step, we train two ADMs: one for normal traffic signs and another for damaged ones. Each model is trained using AdamW optimizer with a learning rate 1×10^{-4} for 300,000 iterations.

The second step in EGSDE involves training the Domain-Specific Extractor (DSE). Although pre-train weights from the guided diffusion are typically used in original research, no existing tasks are similar to normal and damaged traffic sign images. Therefore, we trained the DSE from scratch using the AdamW [28] optimizer with a learning rate 3×10^{-4}. The training process involved two classes of traffic signs and was repeated for 10,000 iterations.

The EGSDE model is utilized to predict noise and generate images. To generate an image using this model, noise must be added to the original image, with a maximum noise level of 1000 (T). The parameter M determines the amount of noise to be added to the image, ranging from 0.3 T to 0.7 T.

C. Dataset

As part of our experiment, we studied traffic sign images using an existing dataset [29]. While analyzing the data, we found that the images in the dataset lack damaged traffic signs, which could compromise the accuracy of our study. To address this issue and obtain more comprehensive data, we collected additional images from various sources on the internet.

The experimental dataset consists of two types of images: normal and damaged. We have divided the data into training and testing sets to avoid overfitting. The training set contains a total of 3,619 images, with 2,919 images labeled as normal and 700 images labeled as damaged. On the other hand, the testing set comprises 600 images, with an equal number of normal and damaged images. Fig. 2 shows example images from our dataset.

Fig. 2. Example image from the dataset.

D. Evaluation Metrics

To evaluate the model's performance, we use the Fréchet Inception Distance (FID) method [30]. This method determines the similarity between two sets of images and is reliable for assessing visual quality. FID is commonly used to evaluate the performance of Generative Adversarial Networks. FID calculates the Fréchet distance

between two Gaussian distributions, which are based on the feature representations of the Inception network [31].

$$\text{FID} = \|\mu_1 - \mu_2\|^2 + \text{Tr}(\sigma_1 + \sigma_2 - 2\sqrt{\sigma_1 \times \sigma_2}) \quad (1)$$

Eq. (1) provides a detailed explanation of how FID is calculated. μ_1 and μ_2 represent the mean of features in real and generated images, while σ_1 and σ_2 are the covariance matrices for the feature vectors of real and generated images, respectively. $\|\mu_1 - \mu_1\|^2$ is the sum squared difference between the two mean vectors, and *Tr* is the trace linear algebra operation. Lower FID scores indicate that the two groups of images have more similar statistics or are similar. A score of 0.0 means that the two groups of images are identical.

IV. Result

This section evaluates the result of translating the image between normal to damaged and damaged to a normal traffic sign. The results of each model are compared using FID as score-based, where lower FID indicates better results. Moreover, we compare and visualize the results to see the differences between each model.

A. FID

Table I presents a comparison of results using FID. The EGSDE has been used to translate the original image by adding different noise levels (M). The parameter M has been set in the range of 0 to T, and the results presented here are for M values of 0.3 T, 0.4 T, 0.5 T, 0.6 T, and 0.7 T. Lower values of M are preferred to preserve the original image structure.

TABLE I. FID Using Different Methods to Translate between Normal and Damaged Traffic Signs (Lower is Better)

Method	Normal-to-Damaged	Damaged-to-Normal
UVCGANv2	174.22	255.29
EGSDE (M = 0.3 T)	***140.84***	161.94
EGSDE (M = 0.4 T)	143.75	172.25
EGSDE (M = 0.5 T)	148.22	184.27
EGSDE (M = 0.6 T)	153.24	182.57
EGSDE (M = 0.7 T)	169.29	***158.33***

In the overview, FID results show that EGSDE performs better than UVCGANv2 in all translation tasks. However, the FID values are still high compared to other I2I translation models with other tasks (E.g., male-to-female tasks have an FID of 40). These results suggest that the I2I translation between normal and damaged traffic signs may be successful.

Regarding the FID, we observed that EGSDE performs better in normal to damaged translation tasks when M is low. However, upon examining the images, we found that most EGSDE results with M = 0.3 T are unsuccessful in translating, as the output remains similar to normal traffic signs. More detailed information is provided in the output image comparison section.

It's important to note that FID results may not reflect the quality of the translated images. Therefore, it's necessary to analyze the inputs and outputs of each method to gain deeper insights.

B. Translated Image Comparison

The comparison results of normal-to-damage translation are presented in Fig. 3. We observed that the EGSDE parameter failed to produce the desired translation even when set to 0.7 T. On closer examination, The EGSDE results bore a striking resemblance to the original image, with only minor differences in the text or numbers, which appeared distorted or unclear. These differences were noticeable and could lead to misinterpretation of the translated information.

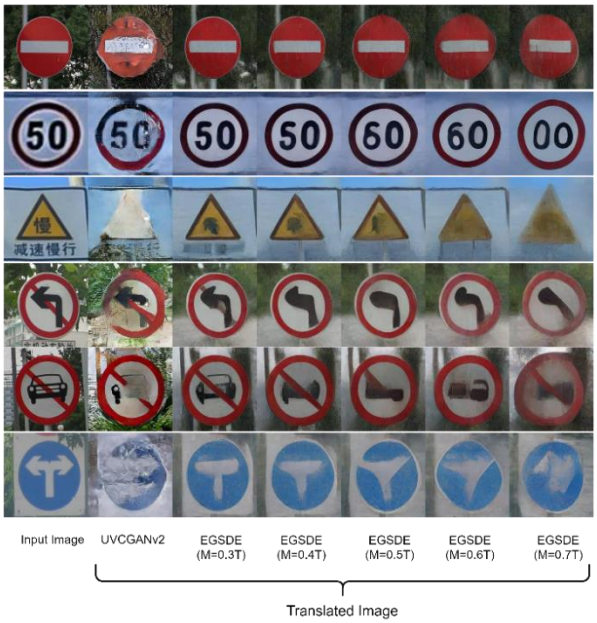

Fig. 3. Translated results for normal-to-damaged.

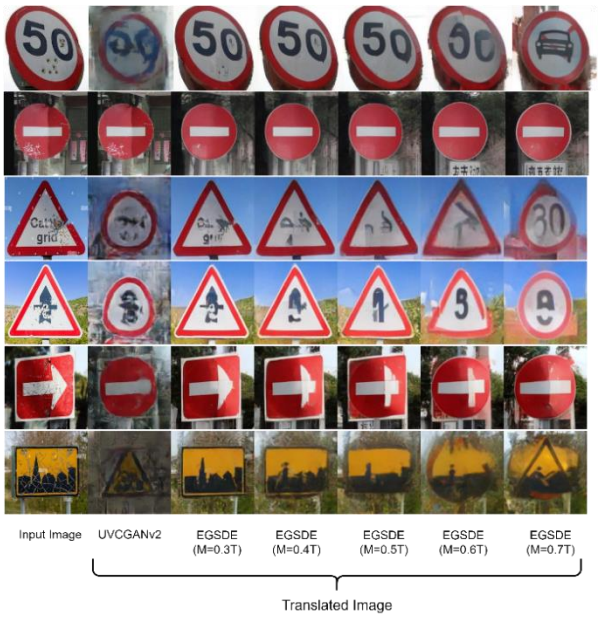

Fig. 4. Translated results for damaged-to-normal.

On the other hand, UVCGANv2 shows potential output, but it still lacks realism. Some traffic signs are also modified to a different shape, such as a circle turning into a triangle. This suggests that UVCGANv2 is not only learning to translate the texture of traffic signs but also modifying their shape, which is different from the objective of this research.

In the analysis of the damage-to-normal translation comparison, Fig. 4 illustrates the outcome. The results obtained from UVCGANv2 exhibit a noticeable failure in translation. The possible reason behind this failure could be that the learning process should have prioritized the texture of the traffic sign specifically but rather focused on the overall texture of the entire image.

EGSDE produces better results than UVCGANv2, especially between 0.3 T and 0.5 T. However, the model can only recover texture, not shape or structure, and realistic results are only possible with undamaged input images.

V. Discussion

Based on our observations, we have noticed that the success or failure of translations relies on the traffic signs' shape, texture, and color. EGSDE with a low M parameter could maintain the input image's structure, whereas UVCGANv2 produced some results that altered the shape of the traffic signs. Fig. 5 illustrates an example of a failed translation and a distorted input traffic sign shape.

Fig. 5. An example of failure translated from normal to damaged using UVCGANv2.

Our assumption for this issue is that UVCGANv2 may mistake the shape of a traffic sign in the input with other shapes present in the training dataset. To avoid translating to the wrong shape issue, we can create a separate class for each traffic sign shape instead of just "normal" and "damaged". However, this would require training UVCGANv2 for each shape, such as a normal circle traffic sign and a damaged circle traffic sign. Another approach to enhance the accuracy of translation is to provide information about the expected output shape in the input by adding a condition.

For translated image texture, the EGSDE method with a low M value allows the texture to remain detailed. On the other hand, if you use a higher M value, the output may not be related to the input image. EGSDE still requires a lower M parameter to translate the traffic sign accurately.

Based on our observations, we have noticed that traffic signs with faded colors often fail to be translated. We assume that our experimental dataset doesn't have enough samples for this kind of damaged traffic sign. An example of this failure to translate is shown in Fig. 6. We suggest finding more samples and experimenting for more insight to avoid this faded-colored translation.

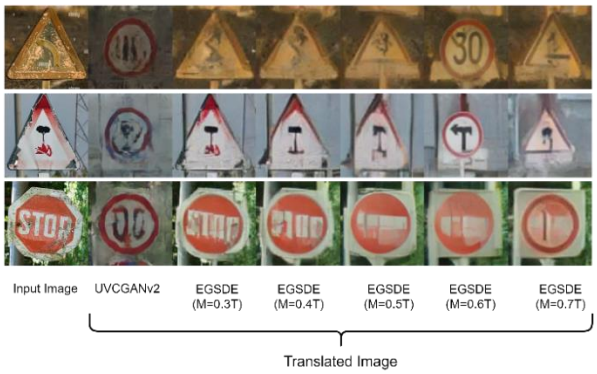

Fig. 6. An example of failure to translate is when the input image has a faded color.

For potential improvement, we recommend two possible solutions. The first is to include the output shape condition for the translate model. The second is to locate additional datasets and categorize them based on various labels rather than just normal and damaged ones.

VI. Conclusion

Our research paper encountered a significant issue: the existing dataset's lack of damaged traffic signs. This issue can be negative impact the performance of autonomous vehicles or driver monitoring systems that rely on object detection because such systems require a large amount of diverse data to train the model effectively. To overcome this issue, we utilized the image-to-image (I2I) translation technique to modify existing images, thereby changing the traffic sign's condition to create a suitable model to increase the quantity of road traffic data with different conditions. We conducted experiments by translating normal and damaged traffic signs using two state-of-the-art unpaired I2I translations, UVCGANv2 and EGSDE. Our experiment revealed that UVCGANv2 effectively translates normal traffic signs to damaged ones, while EGSDE has more potential for translating damaged traffic signs to normal ones.

However, UVCGANv2 and EGSDE are not specifically for translating normal and damaged traffic signs, and both models need improvement to work well with large-scale traffic sign datasets. In future work, we plan to customize the I2I translation for traffic sign tasks by adding conditional factors, such as the shape of the output, to guide the output and design the label. The dataset should

be properly collected and organized into appropriate labels or multilabel, such as (speed sign, normal, old) or (speed sign, damaged, new). Proper labels can help the image generative model gain more insight into the traffic sign.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Rung-Ching Chen conducted the research; Ming-Zhong Fan and William Eric Manongga collected the data for the experiment. Chayanon Sub-r-pa analyzed the data; Chayanon Sub-r-pa and Ming-Zhong Fan wrote the paper; all authors approved the final version.

FUNDING

This paper is supported by the National Science and Technology Council, Taiwan. The Nos are NSTC-112-2221-E-324-003-MY3 and NSTC-112-2221-E-324-011.

REFERENCES

[1] T. Primya, G. Kanagaraj, G. Subashini, R. Divakar, and B. Vishnupriya, "Identification of traffic signs for the prevention of road accidents using convolution neural network," *International Conference on Internet of Things*, pp. 35–44, 2022.

[2] A. Trpković, M. Šelmić, and S. Jevremović, "Model for the identification and classification of partially damaged and vandalized traffic signs," *Springer Science and Business Media LLC*, vol. 25, no. 10, pp. 3953–3965, Jul 2021.

[3] K. Radoš, J. Downes, D.-S. Pham, and A. Krishna, "End-to-end traffic sign damage assessment," in *Proc. 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Nov. 2022.

[4] C. You, C. Wen, H. Luo, C. Wang, and J. Li, "Rapid traffic sign damage inspection in natural scenes using mobile laser scanning data," in *Proc. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Jul 2017, pp. 6271–6274.

[5] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint, arXiv:1511.06434, 2015.

[7] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. International Conference on Machine Learning*, 2017, pp. 214–223.

[8] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell*, 2023.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer "High-resolution image synthesis with latent diffusion models," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.

[10] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Procedia Computer Science* vol. 199, pp. 1066–1073, 2022.

[11] C. Dewi, R. C. Chen, Y. C. Zhuang, X. Jiang, and H. Yu, "Recognizing road surface traffic signs based on YOLO models considering image flips," *Big Data and Cognitive Computing*, vol. 7, no. 1, 54, Mar 2023.

[12] R.-C. Chen, V. S. Saravanarajan, and H.-T. Hung, "Monitoring the behaviours of pet cat based on YOLO model and raspberry Pi," *International Journal of Applied Science and Engineering*, vol. 18, no. 5, pp. 1–12, Sep. 2021.

[13] D. Torbunov *et al.*, "UVCGAN v2: An improved cycle-consistent GAN for unpaired image-to-image translation," arXiv preprint, arXiv:2303.16280, 2023.

[14] M. Zhao, F. Bao, C. Li, and J. Zhu, "Egsde: Unpaired image-to-image translation via energy-guided stochastic differential Eqs," *Advances Neural Information Processing Systems*, vol. 35, pp. 3609–3623, 2022.

[15] U. Khamdamov, M. Umarov, J. Elov, S. Khalilov, and I. Narzullayev, "Uzbek traffic sign dataset for traffic sign detection and recognition systems," in *Proc. 2022 International Conference on Information Science and Communications Technologies (ICISCT)*, Sep. 2022, pp. 1–5.

[16] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[17] G. Zhao, Z. Cai, X. Wang, and X. Dang, "GAN data augmentation methods in rock classification," *Appl. Sci.*, vol. 13, no. 9, 5316, 2023.

[18] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, "Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks," *Sci. Rep.*, vol. 9, no. 1, 16884, 2019.

[19] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[20] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015*, Munich, Germany, 2015, pp. 234–241.

[22] K. Han *et al*., "A Survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2023.

[23] D. Torbunov *et al.*, "Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation," in *Proc. the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 702–712.

[24] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efro, "Context encoders: Feature learning by inpainting," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[25] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint, arXiv:1810.04805, 2018.

[26] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. International Conference on Machine Learning*, 2021 pp. 8162–8171.

[27] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[28] Z. Zhuang, M. Liu, A. Cutkosky, and F. Orabona, "Understanding adamw through proximal methods and scale-freeness," arXiv preprint, arXiv:2202.00089, 2022.

[29] J. Boghean. (January 2023). Damaged Signs Multi-label Computer Vision Project *Roboflow Universe*. [Online]. Available: https://universe.roboflow.com/jayke-boghean-2pxtg/damaged-signs-multi-label

[30] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.