Vehicle Image Translation: Adapting Synthetic Styles to Real-World Scenarios

Qixiang Chen*

Research School of Computer Science
Australian National University
Canberra, Australia

Abstract. This study addresses the challenge of bridging the gap between synthetic and real-world images for vehicle recognition, classification and re-identification, a critical task in the development of intelligent transportation systems. With the increasing complexity of urban traffic and the need for advanced traffic management, the scarcity of real-world training data, compounded by privacy concerns and the labour-intensive nature of data labelling, presents significant hurdles. Our research explores the use of synthetic data for model training, specifically targeting the refinement of synthetic images from the Vehicle-X dataset to closely resemble real-world images from the VeRi dataset. Employing Cycle Generative Adversarial Networks (CycleGAN), we demonstrate the potential of unsupervised domain adaptation to generate realistic, high-quality images for downstream tasks. The study's findings have significant implications for traffic management, route planning, and accident prevention, offering a promising direction for future research in the field.

Keywords: Domain adaptation · Image-to-image translation · Generative Adversarial Networks · Cycle-Consistent

Fig. 1. Results on mapping synthetic images to real-world styles and vice versa. Each pair displays the source and target, with the origin image on the left and the generated image on the right. The top row showcases synthetic images from the Vehicle-X dataset, while the bottom row presents real images from the VeRI dataset.

1 Introduction

In the increasingly crowded urban area and growing demand for intelligent transportation systems, the need for efficient traffic management and safety measures has never been greater, and the ability to accurately recognize and classify vehicles is important. Among the various vehicles on the road, large vehicles such as trucks, buses, and trailers present unique challenges due to their size, maneuverability constraints, and potential safety risks. Recognizing these vehicles in real-time is crucial for traffic management, route planning, and accident prevention.

However, access to real-world training data is a major obstacle. For tasks such as recognition [19], classification [5] and re-identification (ReID) [25,4,13], a fully labelled source domain is essential during the training process. The labelling process is labour-intensive and costly, which is a considerable challenge. Moreover, protecting passenger privacy and low-quality capturing with motion blur adds another layer of complexity. This requires careful selection of images and meticulous masking of recognizable details such as license plates, which further increases the cost.

* The author conducted this work while enrolled as a Bachelor' student at ANU, specifically for the course COMP4660 in 2023.

One potential workaround to this challenge is to utilize synthetic data for model training. Such data has the advantage of controlled attribute generation with infinite productivity. However, when models trained on synthetic datasets such as the Vehicle-X dataset [25] are used in real-world scenarios, significant drawbacks arise and performance tends to degrade, which is largely attributed to biases between datasets [4].

To address this problem, a common strategy is unsupervised domain adaptation. Our research aims to take advantage of controlled, annotated synthetic images and make them applicable to the real world. We endeavour to "translate" these synthetic images so that they more closely resemble real-world photographs. This approach has the dual advantage of providing a steady stream of high-quality images for downstream applications and ensuring that models are trained on images that are close to real-world conditions. Our main objective in this paper is to refine the style of the Vehicle-X dataset to make it more consistent with real-world scenarios. To do this, we target real-world images from the VeRi dataset [13] in conjunction with the cycle generative adversarial network (CycleGAN) [26] which is inspired by the ReID task from Yao et al. [25], the domain adaptation ensures that the synthesized images are not only realistic but also of high quality from synthetic image to real-world style (Figure 1, top).

2 Related Works

Generative Adversarial Networks (GANs) as a remarkable approach in the field of deep learning and generative models. Introduced by Goodfellow et al. [6], GANs operate on the principle of a zero-sum game between a generator and a discriminator. Various enhancements and modifications to the original GAN architecture have been proposed to improve stability, generate higher-quality outputs, and enable diverse applications. Some notable variants include Conditional GANs that generate images [22,15,11], as well as video [21] or 3D data [23] based on their certain conditions or labels. For our study, the adversarial loss can help us discover the potential mapping between domains.

Image-to-Image translation (I2I) focuses on finding the mapping relation from the source domain to the target domain while preserving the content structure. Early proposed variational autoencoder (VAE) [17] with the inspiration of the Helmholtz machine [3] uses an encoder to estimate data distributions and a decoder to map latent variables, optimizing its model to closely match the given data. Recent approaches [11] using GANs to reach a balance between a generator that creates transferred images and a discriminator that identifies the origin versus transferred images. The related application involves tasks such as style transfer [18,26], image segmentation [7], and image colourization [9,24]. Unlike the translation along with the clear paired mapping, the fine-grained classes in the vehicle raise the difficulty of acquiring synthetic to real paired examples.

Unpaired Image-to-Image translation is more applicable for a broader range of applications. Addressing the challenges of unpaired settings, recent methodologies have incorporated techniques like the Bayesian framework [16], cycle-consistency constraints [26], and multi-modal approaches [1]. In [25], the proposed attribute descent algorithm aims to achieve the content domain adaptation (DA) for the re-ID task based on CycleGAN [26], minimizing the discrepancy between domains with Fréchet Inception Distance (FID) [10]. Similar to our study, we emphasize discovering a mapping function that not only bridges the gap between two domains but also retains the essential feature characteristics specific to vehicles.

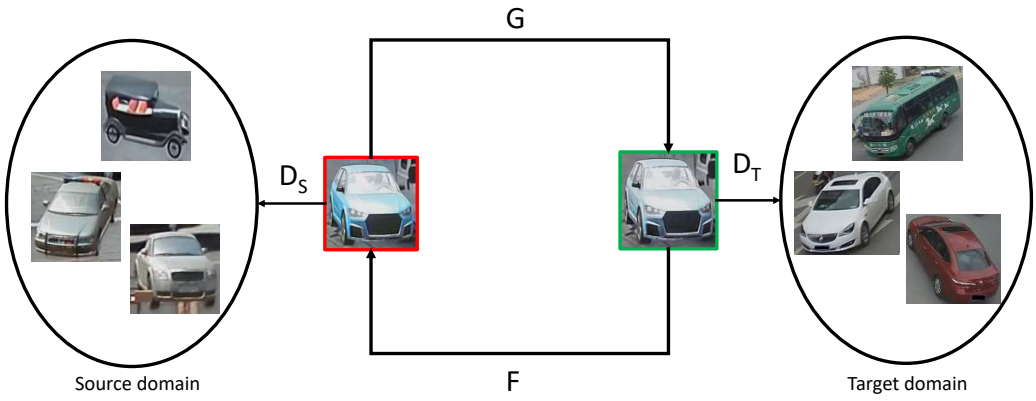

Fig. 2. The CycleGAN Model Framework: This schematic illustrates the dual structure of the model, where $\mathbf{D}_{\mathcal{T}}$ and $\mathbf{G}$ work in concert to refine synthetic images towards the target domain style, while $\mathbf{D}_{\mathcal{S}}$ and $\mathbf{F}$ collaborate to ensure fidelity in the reverse translation.

3 Methdology

In our approach, we utilize a labelled dataset $\mathcal{S}$ from the source domain, which consists of synthetic vehicle images, and an unlabeled dataset $\mathcal{T}$ from the target domain, representing real-world vehicle images. The primary objective is to adeptly translate these synthetic images, ensuring they align closely with the style of real-world imagery while maintaining sufficient characteristics.

3.1 Cycle Generative Adversarial Network (CycleGAN)

Our backbone network in Figure 2 is CycleGAN introduced by Zhu et al. [26]. This model is uniquely designed to handle unpaired image-to-image translation tasks. It comprises two sets of generators and discriminators, $\{\mathbf{G}, \mathbf{D}_\mathcal{T}\}$ and $\{\mathbf{F}, \mathbf{D}_\mathcal{S}\}$. These components work in tandem, the generators mapping images from one domain to the other and vice versa, while the discriminators play a key role in distinguishing between original and translated images, ensuring the generated images are high fidelity and close to the target domain.

3.2 Adversarial Loss

Adversarial Loss ensures that the translated images are indistinguishable from images in the target domain. It pushes the generator to produce images that the discriminator cannot easily differentiate from real images, thereby ensuring the synthetic images are translated to closely resemble real-world styles, the adversarial loss for the mapping function $\mathbf{G} : \mathcal{S} \to \mathcal{T}$ is

$$\mathcal{L}_{\text{GAN}}(\mathbf{G}, \mathbf{D}_\mathcal{T}, p_\mathcal{S}, p_\mathcal{T}) = \mathbb{E}_{y \sim p_\mathcal{T}}[(\mathbf{D}_\mathcal{T}(y))^2] + \mathbb{E}_{x \sim p_\mathcal{S}}[(1 - \mathbf{D}_\mathcal{T}(\mathbf{G}(x)))^2] \tag{1}$$

where $p_\mathcal{S}$ and $p_\mathcal{T}$ denote the sample distributions of two domains and the corresponding adversarial loss for the mapping function $\mathbf{F} : \mathcal{T} \to \mathcal{S}$ is

$$\mathcal{L}_{\text{GAN}}(\mathbf{F}, \mathbf{D}_\mathcal{S}, p_\mathcal{T}, p_\mathcal{S}) = \mathbb{E}_{x \sim p_\mathcal{S}}[(\mathbf{D}_\mathcal{S}(x))^2] + \mathbb{E}_{y \sim p_\mathcal{T}}[(1 - \mathbf{D}_\mathcal{S}(\mathbf{F}(y)))^2] \tag{2}$$

Hence, generators try to map the origin image close to the target domain and discriminators learn to distinguish between origin and translated images, hence the overall object is to $\min_\mathbf{G} \max_{\mathbf{D}_\mathcal{T}} \mathcal{L}_{\text{GAN}}(\mathbf{G}, \mathbf{D}_\mathcal{T}, p_\mathcal{S}, p_\mathcal{T})$ and $\min_\mathbf{F} \max_{\mathbf{D}_\mathcal{S}} \mathcal{L}_{\text{GAN}}(\mathbf{F}, \mathbf{D}_\mathcal{S}, p_\mathcal{T}, p_\mathcal{S})$.

3.3 Cycle Consistency Loss

Cycle Consistency Loss is a crucial component of the CycleGAN model. It ensures that an image, when translated from the source domain to the target domain and then back to the source domain, retains its original characteristics. With the reasonable reduction of the mapping space, this loss ensures that the translation process does not result in any significant loss of information and that the mapping between the two domains is consistent, which satisfies the constraints $\mathbf{F}(\mathbf{G}(x)) \approx x$ and $\mathbf{G}(\mathbf{F}(y)) \approx y$, and the loss is defined as

$$\mathcal{L}_{\text{cyc}}(\mathbf{G}, \mathbf{F}) = \mathbb{E}_{x \sim p_\mathcal{S}} \|\mathbf{F}(\mathbf{G}(x)) - x\|_1 + \mathbb{E}_{y \sim p_\mathcal{T}} \|\mathbf{G}(\mathbf{F}(y)) - y\|_1 \tag{3}$$

3.4 Identity Loss

Identity loss ensures the preservation of key features during the image translation process. If a real image from the target domain is fed into the generator, the output should be the same or close to the original image. Similarly, when a synthetic image from the source domain is provided to the generator, it should ideally remain unchanged after the translation. The identity loss is defined as:

$$\mathcal{L}_{\text{idt}}(\mathbf{G}, \mathbf{F}) = \mathbb{E}_{x \sim p_\mathcal{S}} \|\mathbf{G}(x) - x\|_1 + \mathbb{E}_{y \sim p_\mathcal{T}} \|\mathbf{F}(y) - y\|_1 \tag{4}$$

This results in more coherent and realistic translations, reducing artifacts and ensuring that the translated images maintain the essential features of the original images.

3.5 Full Objective

Our full objective is to integrate both adversarial and cycle consistency losses to ensure high-quality image translation between the source and target domains. This can be represented by the following combined loss function:

$$\begin{aligned}\mathcal{L}(\mathbf{G}, \mathbf{F}, p_\mathcal{S}, p_\mathcal{T}) = \mathcal{L}_{\text{GAN}}(\mathbf{G}, \mathbf{D}_\mathcal{T}, p_\mathcal{S}, p_\mathcal{T}) + \mathcal{L}_{\text{GAN}}(\mathbf{F}, \mathbf{D}_\mathcal{S}, p_\mathcal{T}, p_\mathcal{S}) \\ + \lambda_1 \mathcal{L}_{\text{cyc}}(\mathbf{G}, \mathbf{F}) + \lambda_2 \mathcal{L}_{\text{idt}}(\mathbf{G}, \mathbf{F})\end{aligned}$$

where λ_1 acts as a weighting factor, determining the relative importance of the cycle consistency loss and also λ_2 determining the weights of the identity loss in the combined objective. Adjusting scalars allows for fine-tuning the balance between the adversarial, cycle consistency components and identity components, ensuring optimal translation performance.

Given the inherent challenges of converting synthetic vehicle images to real-world style without pairwise data, the architecture of CycleGAN combined losses is an ideal solution. It allows us to perform pairwise-free translations while preserving key image features making it particularly well-suited to our task.

4 Experiments

4.1 Experiments Settings

Datasets. Our research is focused on the transformation of synthetic vehicle images to real-world environments. The experiments leverage two primary datasets:

Vehicle-X is a large synthetic annotated dataset [25], created in Unity with 1,362 vehicle models. The proposed attribute descent approach seeks to minimize the discrepancy between the gaps using the FID metric. This dataset achieves competitive accuracy in the vehicle ReID application. It consists of $45,438$ training images and $15,142$ testing images.

VeRi-776 is a dataset [14] captured from real-world scenarios as our target domain. It was compiled using 20 different cameras and includes 776 unique vehicles, offering a wide range of variability. The dataset comprises a total of $49,357$ images.

To ensure the integrity of our experiments, we examine both datasets for any redundant or duplicate entries within the training and testing sets. The label distribution in Figure 3, which is crucial for understanding the dataset composition.

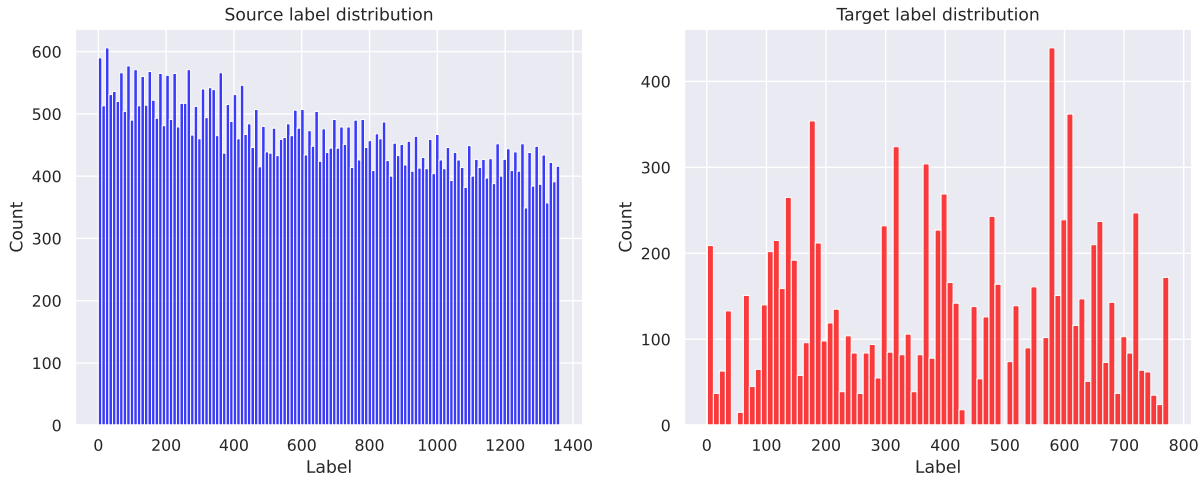

Fig. 3. The distribution of vehicle labels in both the Vehicle-X and VeRi-776 datasets, highlighting the balance and variety essential for robust model training.

Evaluation metric. To assess the performance of our image translation model, we utilize the Fréchet Inception Distance (FID) [10], a metric that quantifies the disparity between the generated image distribution and the distribution of real images. This is achieved by comparing the features extracted by the InceptionV3 model [20] from both sets of images. A lower FID score indicates a smaller distance between the two distributions, signifying better performance of the image translation. The FID score is derived using the following formula:

$$\mathrm{FID} = \|\mu_{\mathrm{real}} - \mu_{\mathrm{generated}}\|^2 + \mathrm{Tr}(\Sigma_{\mathrm{real}} + \Sigma_{\mathrm{generated}} - 2(\Sigma_{\mathrm{real}}\Sigma_{\mathrm{generated}})^{\frac{1}{2}}) \tag{5}$$

where μ_{real} and $\mu_{\mathrm{generated}}$ are the feature-wise means of the real and generated images, respectively, and Σ_{real} and $\Sigma_{\mathrm{generated}}$ are the corresponding covariance matrices. The trace function sums the diagonal elements of the resulting matrix, providing a single scalar that represents the Fréchet distance.

Network Architecture Our model follows the original CycleGAN architecture [26]. The generator is composed of a sequence of layers: it starts with three convolutional layers, followed by several residual blocks [8] that are crucial for learning the identity mapping, then proceeds with two fractionally-strided convolutions with a stride of 1/2, and ends with a final convolutional layer that transforms the feature representations into an RGB image.

The discriminator employs a 70x70 PatchGAN structure [11]. This design enables the discriminator to classify whether each 70x70 patch of overlapping image regions is real or fake. The advantage of using PatchGAN is that it can be applied to images of arbitrary sizes, allowing it to assess the authenticity of local image patches and, by extension, the entire image. This localized approach to discrimination encourages the generator to focus on high-fidelity, detailed translations at the scale of these patches, which collectively contribute to the realism of the entire image.

4.2 Implementation Details

In our implementation, we adopt the training stabilization techniques from the original CycleGAN framework. The codebase is developed by building upon the implementation provided by Yao et al. [25], ensuring consistency with established methodologies.

To mitigate model oscillation in GAN training, we utilize the image buffer that stores up to 50 previously generated images. During training, the discriminators sample from this image pool instead of using the latest generated images directly. This strategy provides the historical knowledgebase of the fake images and smooths the training process by preventing the discriminators from overfitting to the most recent generator outputs.

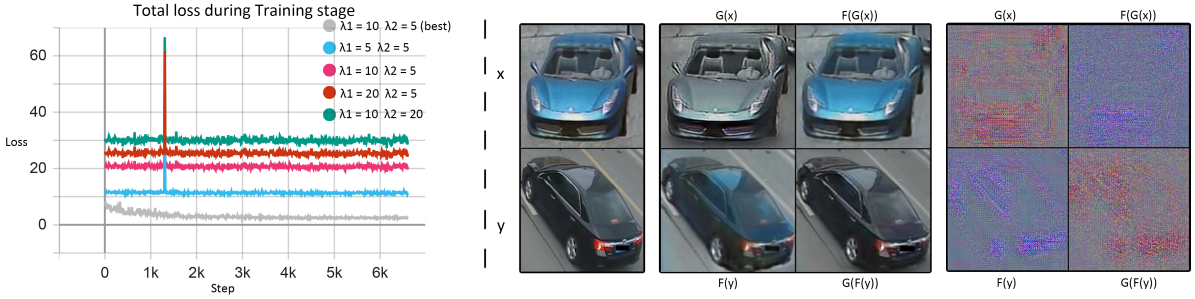

Fig. 4. Comparison of model training outcomes under various λ settings. (*left*) The loss trajectory during training with the optimal hyperparameters ($\lambda_1 = 10$, $\lambda_2 = 5$) shows a significant reduction, while the other configurations do not produce a substantial reduction. (*right*) Training outputs are paired to contrast the effective learning with optimal λ values (on the left) against the stagnant results from other settings, which are generated similar to the result(on the right). This proves the complexities of hyperparameter tuning in GAN models.

Preliminary experiments were conducted to ascertain the optimal hyperparameters. As illustrated in Figure 4, we determined the best-performing hyperparameters to be $\lambda_1 = 10$ and $\lambda_2 = 5$, utilizing the `Adam` optimizer [12] for network updates. In contrast, the use of `SGD` [2] does not result in model convergence.

Fig. 5. Comparison of model training outcomes with two different learning rates and without the scheduler. (*left*) illustrates the loss trajectory during training under the default settings. (*middle*) shows the training output when the learning rate is set to 0.001, revealing the model's inability to learn the appropriate mapping. (*right*) depicts the training output with a learning rate of 0.0001, suggesting that the model learns insufficiently.

The networks are trained from scratch, all starting with an initial learning rate of `0.0002`, as prescribed by the original CycleGAN methodology. The higher and lower learning rates lead to failure, as evidenced by the outcomes depicted in Figure 5.

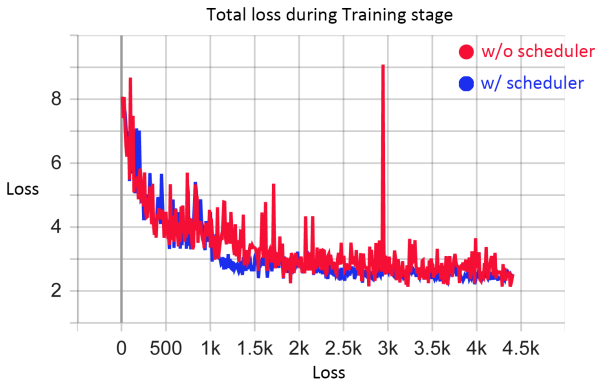

Fig. 6. The graph shows the model's convergence over steps with a fixed learning rate (red) versus a scheduled decay (blue).

Additionally, we evaluated the efficacy of implementing a learning rate scheduler. In this approach, the learning rate is maintained for the first quarter of the total training steps, spanning 8 epochs. Subsequently, the rate is gradually reduced to zero over the following quarter of the training steps. The results in Figure 6, indicate that the model achieves better convergence when trained with the scheduler, this prevents the model from running out of local minimum.

Experiment Results		
Model Index	Model Name	FID ↓
1	lr_0.0002_batch_size_3	118.9905 ± 1.5651
2	schedule_lr_0.0002_batch_size_3	121.3746 ± 1.0153
3	schedule_lr_0.0002_batch_size_12	123.8506 ± 0.7428
4	schedule_lr_0.0002_batch_size_18	123.8545 ± 0.0506
5	lr_0.0002_batch_size_10	123.9388 ± 0.8185
6	schedule_lr_0.0002_batch_size_18_SGD	349.6307 ± 0.0033

Table 1. Summary of experiment results showcasing the impact of different batch sizes, the implementation of a learning rate scheduler, and the choice of optimizer on the FID score. Lower scores indicate better performance, with the best and worst scores highlighted in green and red, respectively.

4.3 Results & Evaluation

The experiments run on the NVIDIA Volta100 GPU, utilizing the PyTorch framework for model implementation and training. The results are analyzed to identify the best combination of hyperparameters that would maximize the performance of the generator responsible for mapping images from the source to the target domain. The model generates the fake image from the test dataset, and evaluates them using the Fréchet Inception Distance (FID) metric to identify the most suitable model for our objectives. Furthermore, a comparative analysis of the actual outcomes from each model was performed in section 4.5.

The results are shown in Table 1, which is calculated based on the mean and standard deviation of the FID scores from the last three epochs of training. It was observed that the Adam optimizer outperformed the SGD optimizer in overall performance. Regarding batch size, a smaller batch size achieved a better FID score. However, it also resulted in a higher standard deviation compared to larger batch sizes. This variability can be attributed to the fine-grained nature of the dataset, which allows for more nuanced updates and potentially more erratic learning paths when smaller batches are used.

As for the implementation of a learning rate scheduler, the results did not conclusively demonstrate an improvement in final performance across all configurations. While the scheduler helped prevent significant loss spikes in models with larger batch sizes, it appeared to restrict the learning potential for models with smaller batch sizes, possibly due to the reduced opportunity for the model to adjust to the data intricacies within the limited learning rate range.

4.4 Discussion

The evaluation results offer insightful revelations about the behaviour of GANs in the context of synthetic-to-real image translation for vehicle datasets. The performance of the Adam optimizer underscores its robustness and adaptability, particularly in scenarios where the optimization landscape is complex and multi-modal. This finding aligns with the existing literature that often chooses Adam over the learning process.

The observation that smaller batch sizes have better FID scores, although with higher variability, suggests that smaller batches allow for more frequent updates. However, the increased standard deviation indicates a trade-off between performance and stability. This trade-off could be a focal point for future research that could offer better performance.

The learning rate scheduler's impact on smaller batch sizes raises questions about the balance between exploration and exploitation in the training of GANs. It may be beneficial to explore other scheduling methods that can fit with the model.

Moreover, the fine-grained nature of the dataset and its impact on model training warrants further examination. Understanding how different data attributes and their representations affect training could lead to more targeted data augmentation strategies or the design of more specialized network architectures that are better suited to handle such datasets.

4.5 Discussion on Visual Outcomes

While the FID scores provide a quantitative measure of the model's performance, a closer inspection of the visual outcomes reveals additional insights. The slight differences in FID scores are not always perceptible in the actual outputs. Models are adept at converting lighter, more vibrant colouring into the more subdued tones typical of real-world captures. This adaptation is a positive step towards realistic image translation. However, a visual assessment indicates that the translation, although heading in the right direction, still falls short of the desired level of realism. The generated images require further refinement to achieve a more precise and indistinguishable representation of real vehicle images.

One possible explanation for the shortfall in translation quality could be the fine-grained nature of the vehicle classes within the datasets. The model may struggle to learn from specific classes due to their subtle and intricate variations. Moreover, while the overall style appears to be captured effectively, the nuanced details that contribute to a vehicle's class identity may be lost or inadequately represented in the translation process. This observation suggests that future work could benefit from a more granular approach to style transfer, perhaps by incorporating class-specific style adaptation mechanisms or by enhancing the model's capacity to retain class-defining features during translation.

The current findings underscore the complexity of image-to-image translation tasks, especially when dealing with fine-grained categories. As such, future research should consider not only the global style transfer but also the preservation and accurate rendering of class-specific attributes. This dual focus could lead to more sophisticated generative models that excel in both style adaptation and class representation, thereby pushing the boundaries of synthetic-to-real image translation.

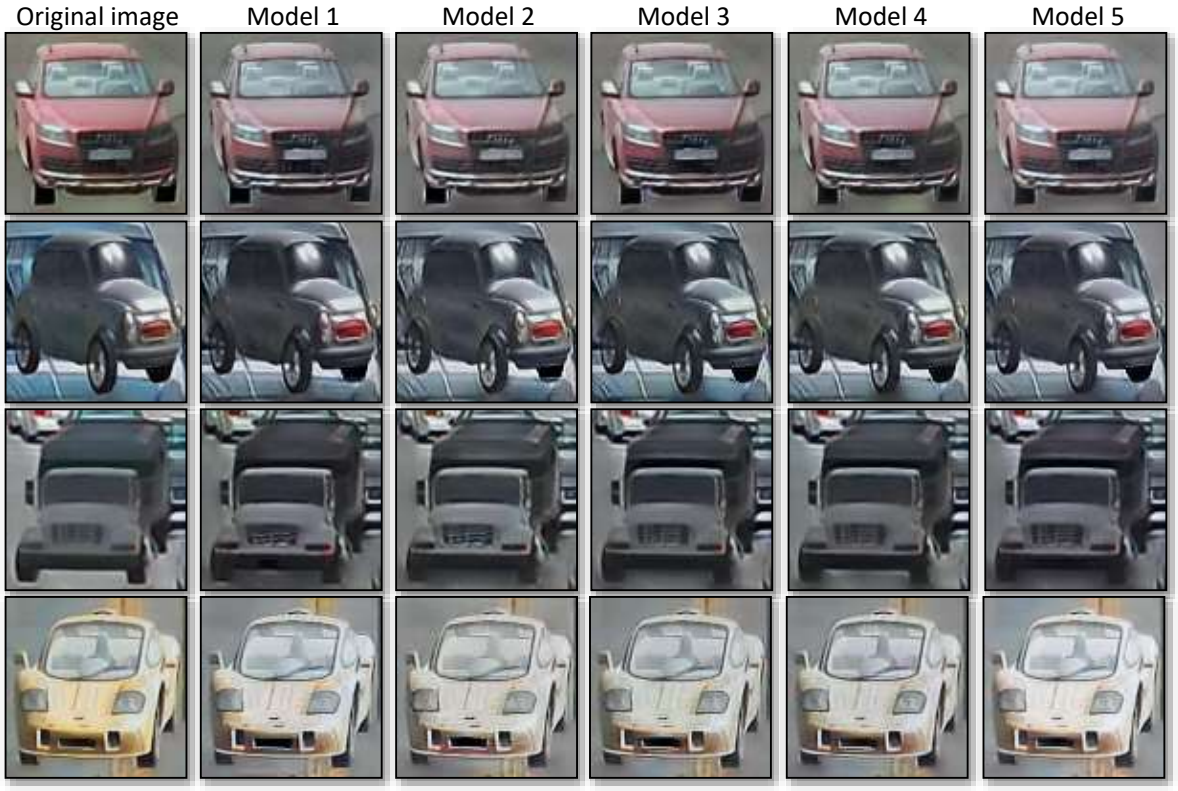

Fig. 7. The test outcomes of the first five models, the leftmost is the original image.

5 Conclusions

This study aims to bridge the gap between synthetic and real-world images for vehicle recognition, classification and re-identification, a critical component in the advancement of intelligent transportation systems. Our main findings indicate that through the use of Cycle Generative Adversarial Networks (CycleGAN), we can effectively translate synthetic vehicle images to closely resemble real-world scenarios. The significance of this work lies in its potential to alleviate the challenges associated with the scarcity of labelled real-world data, a hurdle that often impedes the progress of machine learning models in practical applications.

The implications of our study are far-reaching. By demonstrating that synthetic data can be adapted to mimic real-world conditions, we pave the way for more robust vehicle classification systems. These systems are essential for enhancing traffic management, route planning, and safety measures, particularly in urban environments where the recognition of large vehicles is crucial.

However, the variability in the performance of the model, as indicated by the standard deviation in the FID scores, suggests that there is room for improvement in consistency with better performance. Additionally, the visual outcomes of the translated images still require refinement to fully capture the nuances of real-world images.

For future work, we suggest exploring further architectures and training methodologies that could further reduce the gap between synthetic and real-world image distributions. Investigating the impact of different data attributes on the training process could lead to more targeted data augmentation strategies or the development of specialized network architectures.

In conclusion, our study contributes a significant step towards the utilization of synthetic data for real-world applications in vehicle classification. It opens new avenues for research and development in the field of intelligent transportation systems, with the potential to make a substantial impact on traffic safety and management.

6 Acknowledgement

The author would like to express their gratitude to ChatGPT for its assistance in refining and polishing the content of this paper.

References

1. Almahairi, A., Rajeshwar, S., Sordoni, A., Bachman, P., Courville, A.: Augmented cyclegan: Learning many-to-many mappings from unpaired data. In: International conference on machine learning. pp. 195–204. PMLR (2018)
2. Bottou, L., et al.: Stochastic gradient learning in neural networks. Proceedings of Neuro-Nımes **91**(8), 12 (1991)
3. Dayan, P., Hinton, G.E., Neal, R.M., Zemel, R.S.: The helmholtz machine. Neural computation **7**(5), 889–904 (1995)
4. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 994–1003 (2018)
5. Duarte, M.F., Hu, Y.H.: Vehicle classification in distributed sensor networks. Journal of Parallel and Distributed Computing **64**(7), 826–838 (2004)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
7. Guo, X., Wang, Z., Yang, Q., Lv, W., Liu, X., Wu, Q., Huang, J.: Gan-based virtual-to-real image translation for urban scene semantic segmentation. Neurocomputing **394**, 127–135 (2020)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. He, M., Chen, D., Liao, J., Sander, P.V., Yuan, L.: Deep exemplar-based colorization. ACM Transactions on Graphics (TOG) **37**(4), 1–16 (2018)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Liu, X., Liu, W., Mei, T., Ma, H.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 869–884. Springer (2016)
14. Liu, X., Liu, W., Mei, T., Ma, H.: Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Transactions on Multimedia **20**(3), 645–658 (2017)
15. Perarnau, G., Van De Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible conditional gans for image editing. arXiv preprint arXiv:1611.06355 (2016)
16. Resales, Achan, Frey: Unsupervised image translation. In: Proceedings Ninth IEEE International Conference on Computer Vision. pp. 472–478. IEEE (2003)
17. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: International conference on machine learning. pp. 1278–1286. PMLR (2014)
18. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: A stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2287–2296 (June 2021)
19. Sivaraman, S., Trivedi, M.M.: A general active-learning framework for on-road vehicle recognition and tracking. IEEE Transactions on intelligent transportation systems **11**(2), 267–276 (2010)
20. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
21. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. Advances in neural information processing systems **29** (2016)
22. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
23. Wu, J., Zhang, C., Xue, T., Freeman, B., Tenenbaum, J.: Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. Advances in neural information processing systems **29** (2016)
24. Xu, Z., Wang, T., Fang, F., Sheng, Y., Zhang, G.: Stylization-based architecture for fast deep exemplar colorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9363–9372 (2020)
25. Yao, Y., Zheng, L., Yang, X., Naphade, M., Gedeon, T.: Simulating content consistent vehicle datasets with attribute descent. In: ECCV (2020)
26. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)