Adapting Synthetic Eyes: A Study of Pixel and Feature-Level UDA for Traffic Object Detection

André N. Medeiros*, Renato M. Silva†, Jurandy Almeida*, Tiago A. Almeida*
*Intelligent Systems and Data Science Lab (LaSID)
Federal University of São Carlos (UFSCar), Sorocaba, São Paulo, Brazil
Email: andre.medeiros@estudante.ufscar.br, jurandy.almeida@ufscar.br, talmeida@ufscar.br
†Institute of Mathematics and Computer Sciences (ICMC)
University of São Paulo (USP), São Carlos, São Paulo, Brazil
Email: renatoms@icmc.usp.br

***Abstract*—Training robust object detectors for autonomous driving requires vast amounts of annotated data, making the use of synthetic datasets an attractive alternative. However, models trained on synthetic data suffer from a significant performance drop when deployed in the real world due to the "sim-to-real" domain gap. Unsupervised Domain Adaptation (UDA) aims to solve this problem without requiring expensive target domain annotations. This paper conducts an empirical study comparing two leading but philosophically different UDA paradigms: pixel-level adaptation via image-to-image translation (CycleGAN) and feature-level adaptation via confidence-aware data mixing (ConfMix). We evaluate these methods on challenging synthetic-to-real adaptation tasks, using DOLPHINS and SIM10K as source domains, and Cityscapes and nuScenes as target domains. Our findings demonstrate that the feature-level mixing strategy of ConfMix provides more significant and robust performance gains than pixel-level translation with CycleGAN. Furthermore, we introduce and evaluate a hybrid method, TransConfMix, which yields mixed results, highlighting the complexities of combining these techniques. Our work provides clear evidence and practical guidance on the effectiveness of different UDA strategies, concluding that directly adapting the model's learning process is a more potent approach than preprocessing the data for this critical application.**

I. Introduction

The perception systems of autonomous vehicles rely on the accurate detection of traffic elements such as vehicles and pedestrians to ensure safety and operational efficiency [1]. Modern deep learning object detectors have demonstrated remarkable performance, driven by complex architectures and, crucially, the availability of vast, meticulously annotated datasets [2]–[5]. However, the manual annotation of real-world data is a prohibitively expensive and time-consuming bottleneck. This has spurred immense interest in using realistic synthetic data, which offers a scalable and cost-effective alternative for training robust models. Furthermore, simulation provides a safe environment to generate data for rare or high-risk scenarios that are impractical to capture in the real world [6].

Despite the promise of synthetic data, a fundamental challenge persists: the domain gap. Models trained exclusively on simulated data exhibit a significant performance degradation when deployed in the real world. This gap arises from discrepancies in visual style (e.g., texture, lighting, sensor noise) and content distribution (e.g., object density, scene layout) between the synthetic source domain and the real-world target domain. To harness the full potential of synthetic data, it is imperative to bridge this sim-to-real gap.

Unsupervised Domain Adaptation (UDA) offers a compelling solution, seeking to adapt a model trained on a labeled source domain to an unlabeled target domain. In the context of object detection, UDA approaches have diverged into two main categories. The first is pixel-level adaptation, which uses generative models to transform the appearance of source images to mimic the target domain's style. This approach is highly interpretable, as the quality of the adaptation can be visually inspected before detector training. The second category focuses on feature-level adaptation, which aligns feature distributions or, in more recent methods, employs sophisticated data-mixing strategies to teach the detector robustness directly. These methods adapt the model itself rather than the input data.

This paper presents an empirical study of these two competing UDA philosophies for the critical task of adapting synthetic object detectors to real-world traffic scenes. We ask: Is it more effective to change the appearance of the data (pixel-level) or to change how the model learns from disparate data (feature-level mixing)? To answer this, we investigate two representative, state-of-the-art methods:

- CycleGAN [7], a canonical method for pixel-level adaptation via unpaired image-to-image translation.
- ConfMix [8], an advanced technique that performs confidence-aware mixing of source and target images and labels.

By evaluating these methods on challenging synthetic-to-real benchmarks, our study provides the following contributions:

- We conduct a head-to-head comparison of pixel-level and data-mixing UDA strategies, providing a clear analysis of their effectiveness in bridging the sim-to-real gap for object detection.
- We analyze the distinct advantages and failure modes of each methodology, offering insights into their practical utility and robustness in complex driving scenarios.

- We establish strong benchmark results on public datasets, comparing against source-only (lower bound) and fully-supervised (upper bound) models to contextualize the performance of these UDA techniques.

II. Related Work

Our work is situated at the intersection of object detection and unsupervised domain adaptation. We review the key developments in these areas, focusing on the two major UDA paradigms relevant to our study.

A. Unsupervised Domain Adaptation for Object Detection

The challenge of domain shift in object detection, especially in the sim-to-real context, has attracted significant research interest [9], [10]. Early UDA methods for detection focused on aligning feature distributions between source and target domains at either the image or instance level, often using adversarial training to encourage the model to learn domain-invariant representations. These methods operate directly within the detector's architecture to minimize a domain discrepancy metric. While effective, they can add significant complexity to the training pipeline.

B. Pixel-Level Adaptation via Image-to-Image Translation

A conceptually distinct approach is to resolve the domain shift at the data level before training the detector. This paradigm, known as pixel-level adaptation, uses image-to-image translation models to "re-style" source domain images to match the appearance of the target domain. The most influential method in this area is CycleGAN [7], which leverages a cycle-consistency loss to learn translations between two domains without requiring paired images. This is ideal for the UDA setting where a direct correspondence between a synthetic scene and a real-world one does not exist.

The success of this approach has been demonstrated in various synthetic-to-real benchmarks [11]. In the context of autonomous driving, CycleGAN has been used to translate synthetic training data to look more like real-world datasets, thereby improving the performance of downstream detectors [2]. However, the translation process is not without its challenges. Generative models can sometimes introduce visual artifacts or fail to properly translate fine-grained details, particularly in complex traffic scenes with heavy occlusion and a high density of objects [12]. Despite these limitations, pixel-level adaptation offers a powerful and interpretable framework for UDA, as the quality of the adapted dataset can be directly inspected.

C. Adaptation via Pseudo-Labeling and Data Mixing

An alternative to pixel-level translation is to adapt the model through self-training, which typically involves generating pseudo-labels on the unlabeled target data. The model is then retrained on a combination of labeled source data and pseudo-labeled target data. The primary challenge in this approach is ensuring the quality of the pseudo-labels, as noisy or incorrect labels can lead to error accumulation and performance degradation.

To address this, recent methods have proposed more sophisticated data mixing strategies. ConfMix [8] stands out as a state-of-the-art technique in this category. Instead of translating images, ConfMix operates by creating a new training sample by mixing a source image with patches from a target image. Crucially, the patches selected from the target domain are those where the model has the most confident pseudo-detections. This confidence-aware mechanism ensures that the model learns from the most reliable parts of the target domain, progressively adapting its features while mitigating the risk of noise from low-quality pseudo-labels.

D. Position of Our Work

Image translation and data mixing represent two complementary strategies for reducing the domain gap in VRU detection tasks and offer distinct advantages. However, the comparative performance of methods representing these strategies (e.g., CycleGAN and ConfMix) in VRU detection remains an open question. CycleGAN excels in pixel-level adaptation, allowing for highly realistic transformations of the source domain into the style of the target domain. This allows models to better generalize to real-world data. However, CycleGAN is computationally expensive and sensitive to domain-specific details, making it less suitable for real-time applications.

In contrast, ConfMix offers a simple and more efficient approach that progressively adapts object detectors by making use of pseudo-labels with confidence based on the target domain. While this method is computationally cheaper and easier to implement, it may not be as effective in scenarios where high-quality visual transformation is required. The trade-off lies on the ability of CycleGAN to generate more realistic images but at the cost of increased computational complexity, while ConfMix simplifies the adaptation process but may not achieve the same level of fidelity.

In this study, we evaluate the strengths and weaknesses of both approaches in real-world traffic environments. Specifically, we investigate how each approach adapts object detectors trained on simulated traffic data to real-world datasets, such as Cityscapes and NuScenes. The results aim to provide a comprehensive understanding of the practical utility of each method in challenging real-world scenarios.

III. Methodology

Our methodology is designed to provide a clear and direct comparison between pixel-level, feature-level, and hybrid adaptation strategies. We first define our baseline object detector and then detail the three UDA approaches evaluated in this study. Let $S = \{(x_s, y_s)\}$ be the labeled source domain (synthetic data) and $T = \{x_t\}$ be the unlabeled target domain (real-world data). All experiments focus on a single class: *car*.

A. Preliminaries: Baseline Object Detector

For all experiments, we employ the YOLOv5 architecture. YOLOv5 is a highly efficient and powerful one-stage object detector, making it a relevant choice for applications like autonomous driving. The other reason we adopted YOLOv5

is to ensure a fair and direct comparison across all adaptation methods, since it is the detector architecture used in the official ConfMix implementation. The model is trained to minimize a multi-component loss function that includes classification, localization, and objectness scores. When this model is trained only on the source domain S and evaluated on the target domain T, it serves as our "source-only" baseline, establishing the lower bound for performance.

B. Pixel-Level Adaptation via CycleGAN

To adapt at the pixel level, we use CycleGAN [7] to translate the visual style of the source domain to mimic the target domain. This unpaired image-to-image translation framework learns a mapping $G_{S\to T}$ without corresponding image pairs. It consists of two generators, $G_{S\to T}$ and $G_{T\to S}$, and two discriminators, D_T and D_S. The framework is trained with a combination of an adversarial loss and a cycle-consistency loss.

The adversarial loss pushes the generator to create images that are indistinguishable from real images in the target domain. For the mapping $G_{S\to T}$, the loss is:

$$L_{GAN}(G_{S\to T}, D_T) = \mathbb{E}_{x_t\sim T}[\log D_T(x_t)] + \mathbb{E}_{x_s\sim S}[\log(1 - D_T(G_{S\to T}(x_s)))] \quad (1)$$

The cycle-consistency loss ensures that the content of the image is preserved during translation by enforcing that an image translated to the other domain and back should recover the original image. The loss is defined as:

$$L_{cyc}(G_{S\to T}, G_{T\to S}) = \mathbb{E}_{x_s\sim S}[\|G_{T\to S}(G_{S\to T}(x_s)) - x_s\|_1] + \mathbb{E}_{x_t\sim T}[\|G_{S\to T}(G_{T\to S}(x_t)) - x_t\|_1] \quad (2)$$

The full objective function, which includes a symmetric mapping from T to S, is:

$$L_{CycleGAN} = L_{GAN}(G_{S\to T}, D_T) + L_{GAN}(G_{T\to S}, D_S) + \lambda L_{cyc} \quad (3)$$

where λ is a hyperparameter that balances the importance of the losses. After training, we generate a new, translated source dataset $S' = \{(G_{S\to T}(x_s), y_s)\}$. Then, an object detector approach can be trained from scratch on S'.

C. Feature-Level Adaptation via ConfMix

For feature-level adaptation, we employ ConfMix [8], a curriculum-based strategy that mixes source and target data based on model confidence. The process involves several steps:

1) Pseudo-Label Generation: The detector generates pseudo-labels on unlabeled target images x_t. To capture localization uncertainty, the detector is modified to predict the parameters of a Gaussian distribution for each bounding box coordinate:

$$\hat{b} = [\mu_{bx}, \mu_{by}, \mu_{bh}, \mu_{bw}, \Sigma_{bx}, \Sigma_{by}, \Sigma_{bh}, \Sigma_{bw}] \quad (4)$$

where μ and Σ are the predicted means and variances.

2) Confidence Calculation: A localization confidence score, C_{bbx}, is derived from the predicted variances:

$$C_{bbx} = 1 - \text{mean}(\hat{\Sigma}) \quad (5)$$

This is combined with the standard classification confidence, C_{det}, to produce a final confidence score for each pseudo-detection:

$$C_{comb} = C_{det} \cdot C_{bbx} \quad (6)$$

3) Confidence-Aware Mixing: The target image patch with the highest C_{comb} is pasted onto a source image x_s, creating a mixed image x_M. The labels for this mixed image, y_M, combine the source labels and the high-confidence target pseudo-label.
4) Consistency Training: The mixed image x_M is fed through the detector. A consistency loss is applied, enforcing that the detector's output on the mixed image aligns with the combined labels y_M. This loss is the standard detection loss, L_{det}, applied to the mixed sample:

$$L_{cons} = L_{det}(f(x_M), y_M) \quad (7)$$

where f is the detector. The total training loss is a weighted sum of the supervised loss on the source image and the consistency loss on the mixed image:

$$L_{total} = L_{det}(f(x_s), y_s) + w_{cons}L_{cons} \quad (8)$$

D. A Hybrid Approach: TransConfMix

We also propose and evaluate a hybrid strategy, termed TransConfMix. This approach combines pixel-level and feature-level adaptation. The training protocol is identical to ConfMix, but during the mixing step, the source images x_s are replaced by their CycleGAN-translated counterparts from the dataset S'. This aims to minimize the domain gap before mixing, potentially enabling more effective learning.

IV. Experiments

This section details the empirical evaluation of the UDA strategies described in our methodology.

A. Datasets

We use two synthetic datasets as our source domains and two real-world datasets as our target domains.

- Source (Synthetic):
  - DOLPHINS[1] [13]: A dataset of synthetic vehicle-centric road scenes (1920×1080 pixels).
  - SIM10K[2] [14]: A dataset containing 10,000 images with 58,701 annotated cars, generated from the Grand Theft Auto V video game (1920×1080 pixels).
- Target (Real-World):

[1]DOLPHINS: Available at https://dolphins-dataset.net/. Accessed on September 11, 2025.

[2]SIM10K: Available at https://fcav.engin.umich.edu/projects/driving-in-the-matrix. Accessed on September 11, 2025.

– Cityscapes[3] [15]: A large-scale dataset of real-world urban street scenes. We use the vehicle-perspective images (2048×1024 pixels).
– nuScenes[4] [16]: A comprehensive autonomous driving dataset. We use the front-facing camera data (`CAM_FRONT`), which captures diverse environments (1600×900 pixels).

B. Experimental Protocol and Metrics

To provide a comprehensive evaluation, we establish clear performance bounds and follow specific training protocols for each method.

1) Performance Bounds: We define two key performance benchmarks to provide a clear reference for evaluating the UDA strategies:

- Lower Bound (Source-Only): A YOLOv5 model trained for 100 epochs exclusively on a source dataset and evaluated directly on a target dataset.
- Upper Bound (Oracle): A YOLOv5 model trained for 100 epochs and evaluated on the same target dataset. This represents the ideal fully-supervised performance.

2) Adaptation Protocols: To evaluate the UDA strategies, we employed the following methods:

- CycleGAN: Following preprocessing, a CycleGAN model is trained for each source-target pair (e.g., DOLPHINS → Cityscapes). The entire source dataset is then translated. Finally, a YOLOv5 detector is trained for 100 epochs on this translated dataset.
- ConfMix: We follow the original two-stage protocol. First, a YOLOv5 model is trained for 50 epochs on the source data. Second, this model is used to initiate the adaptation stage, training for an additional 50 epochs using the confidence-based mixing strategy.
- TransConfMix: The protocol is identical to ConfMix, but in the second 50-epoch adaptation stage, the source images are replaced by their CycleGAN-translated versions.

3) Evaluation Metrics: We evaluate all models on the target domain's test set, focusing on detecting the car class. Performance is measured using standard COCO-style metrics:

- mAP@0.5: Mean Average Precision at an IoU threshold of 0.5.
- mAP@[0.5:0.95]: The primary COCO metric, averaging mAP over IoU thresholds from 0.5 to 0.95.

C. Implementation Details

In the following, we present the specific details of how we implemented each method:

- Data splits and leakage prevention: For Cityscapes and nuScenes, only the official `train` split was used for pseudo-labeling/adaptation, while the official `val` split was reserved exclusively for evaluation. In nuScenes, we used only `CAM_FRONT` frames and enforced scene-level disjointness. For datasets without official validation/test partitions (e.g., DOLPHINS, SIM10K), we defined fixed, disjoint subsets at the start of the study and kept them immutable throughout all experiments.
- CycleGAN: We use the official PyTorch implementation[5]. Images are resized to a load size of 1024 and then cropped to 512 for training. We train for 100 epochs (the first 50 with a fixed learning rate of 0.0001 and the last 50 with linear decay). We set $\lambda_{cyc} = 10$ and $\lambda_{identity} = 1.0$.
- ConfMix: We use the official implementation[6] with the default hyperparameters provided for YOLOv5.
- YOLOv5: All detector training uses the default configurations from the YOLOv5 repository. The model checkpoint from the final epoch is used for evaluation.

V. Results and Analysis

In this section, we present and analyze the results. We first report the quantitative performance of each UDA method and then provide a qualitative analysis through visualizations of the adaptation process and final detection outputs.

A. Quantitative Results

The primary results of our experiments are summarized in Table I. The table compares the performance of the Source-Only baseline against the three adaptation strategies (CycleGAN, ConfMix, and TransConfMix) on two distinct synthetic-to-real adaptation tasks: DOLPHINS/SIM10K → Cityscapes and DOLPHINS/SIM10K → nuScenes. Performance is measured by mAP@0.5 and mAP@[0.5:0.95].

TABLE I: Results for object detection in target domains (Cityscapes and NuScenes). All values are mAP. The best adaptation method for each metric is in **bold**. The Oracle provides an upper-bound performance reference.

	Method	Cityscapes		NuScenes	
		mAP@0.5	**mAP@[0.5:0.95]**	**mAP@0.5**	**mAP@[0.5:0.95]**
Source: **DOLPHINS**	Source-only	0.1053	0.0481	0.2873	0.1518
	CycleGAN	0.2199	0.1024	0.2619	0.1382
	ConfMix	**0.2883**	**0.1413**	0.3766	**0.1848**
	TransConfMix	0.2300	0.1064	**0.3806**	0.1829
	Oracle	0.5824	0.3817	0.5215	0.3046
Source: **SIM10K**	Source-only	0.3172	0.1998	0.3573	0.1459
	CycleGAN	0.3525	0.2105	0.3766	0.1538
	ConfMix	**0.4082**	**0.2396**	**0.4126**	0.1619
	TransConfMix	0.3725	0.2144	0.4095	**0.1641**
	Oracle	0.5824	0.3817	0.5215	0.3046

[3]Cityscapes: Available at https://www.cityscapes-dataset.com/. Accessed on September 11, 2025.

[4]nuScenes: Available at https://www.nuscenes.org/. Accessed on September 11, 2025.

[5]CycleGAN official PyTorch implementation. Available at https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix. Accessed on September 11, 2025

[6]ConfMix official implementation. Available at https://github.com/giuliomattolin/ConfMix. Accessed on September 11, 2025

B. Analysis of Results

UDA methods successfully bridge the domain gap. As expected, the source-only models perform poorly on both target domains, confirming the existence of a significant domain shift. For instance, the DOLPHINS-trained model achieves only 0.0481 mAP@[0.5:0.95] on Cityscapes. All adaptation methods provide a substantial boost over this baseline. In the DOLPHINS $\rightarrow$ Cityscapes task, ConfMix improves the mAP@[0.5:0.95] to 0.1413, a relative increase of over 190%. This clearly demonstrates that UDA is a critical component for deploying detectors trained on synthetic data.

Feature-level mixing outperforms pixel-level translation. In our comparison, ConfMix consistently outperforms CycleGAN across nearly all experiments. When adapting from SIM10K to Cityscapes, ConfMix achieves a mAP@0.5 of 0.4082, surpassing CycleGAN's 0.3525. An interesting exception is the DOLPHINS $\rightarrow$ NuScenes task, where CycleGAN surprisingly degrades performance compared to the source-only baseline. This suggests that flawed pixel-level translation can be actively harmful, potentially introducing artifacts that confuse the detector more than the original domain gap. Overall, the evidence suggests that directly adapting the model's features via data mixing is a more robust and effective strategy than relying solely on image-style translation.

The hybrid approach shows mixed results. Our proposed TransConfMix method, which combines pixel-translation with feature-mixing, yields intriguing but inconsistent results. In the DOLPHINS $\rightarrow$ Cityscapes task, it underperforms standard ConfMix. This suggests that the initial CycleGAN translation, even if visually plausible, may introduce subtle artifacts that disrupt the delicate confidence-based mixing mechanism of ConfMix. However, in the DOLPHINS $\rightarrow$ NuScenes task, TransConfMix achieves the highest mAP@0.5 score (0.3806), slightly edging out standard ConfMix. This indicates that for certain domain pairs, a preliminary style alignment may provide a slightly better foundation for the feature-mixing stage. The effectiveness of this hybrid approach appears to be highly dependent on the specific source-target pair and the quality of the initial image translation.

C. Qualitative Analysis

To provide qualitative insights, we visualize the adaptation process and final detection results.

Figure 1 illustrates the core of the pixel-level adaptation. The synthetic DOLPHINS image (a) is translated into the style of Cityscapes (b). While the color palette and lighting are successfully adapted, some fine details and textures are altered, which may contribute to the performance gap relative to feature-level methods.

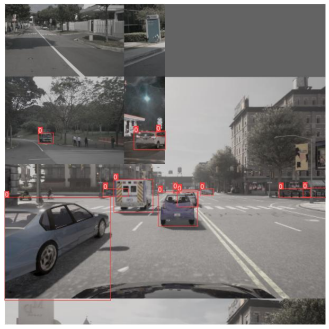
Figure 2 showcases the mechanism of ConfMix. High-confidence regions from the real-world target domains (Cityscapes and NuScenes) are pasted onto synthetic source images. This directly exposes the detector to real-world object appearances and contexts during training, which is likely a key reason for its strong performance.

(a) Original DOLPHINS image.

(b) CycleGAN translation to Cityscapes style.

Fig. 1: An example of CycleGAN's image-to-image translation from the synthetic DOLPHINS domain to the Cityscapes visual style.

(a) DOLPHINS + Cityscapes



(b) DOLPHINS + NuScenes

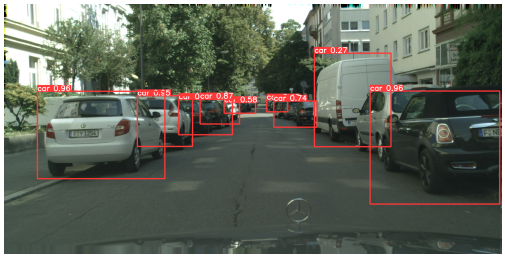
Fig. 2: Region mixing examples for ConfMix when adapting from the DOLPHINS source domain to Cityscapes and NuScenes target domains.

Finally, Figure 3 presents sample detection outputs on the target domains from one of our best-performing adapted models. The model successfully identifies vehicles in varied lighting and traffic conditions, qualitatively demonstrating the effectiveness of the adaptation process.

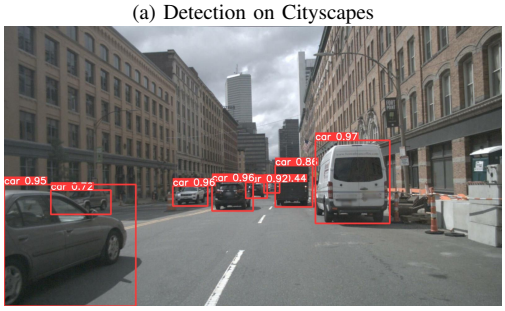
VI. Conclusion

In this paper, we presented a comparative study on UDA for the critical task of vehicle detection, focusing on bridging the domain gap between synthetic source and real-world target domains. We conducted a head-to-head evaluation of a pixel-level adaptation method, CycleGAN, and a feature-level data-mixing method, ConfMix, in addition to a hybrid approach we termed TransConfMix.

Our quantitative and qualitative results lead to a clear conclusion: for the task of sim-to-real object detection, adapting the model through feature-level data mixing is a more effective and robust strategy than adapting the data through pixel-level translation. ConfMix consistently outperformed CycleGAN across our experimental setups, demonstrating that directly exposing the detector to confident pseudo-labels from the target domain within a mixed-data curriculum is highly effective. While visually compelling, the image-to-image translation performed by CycleGAN did not always translate into superior detection accuracy and, in one case, was even detrimental to performance. This suggests that subtle artifacts or imperfect



(a) Detection on Cityscapes



(b) Detection on NuScenes

Fig. 3: Example object detection results on the Cityscapes and NuScenes test sets using a model adapted with ConfMix.

style translations can introduce a new form of "noise" that hinders the detector's learning process.

The findings reinforce the significant value of synthetic data in training perception systems. With effective UDA, models can achieve respectable performance in real-world conditions without any manual annotation in the target domain, dramatically reducing development costs and effort. This has direct implications for improving the safety and reliability of autonomous systems, including the robust detection of vehicles to protect vulnerable road users.

Nonetheless, our study has limitations. The analysis was restricted to a single object class, and the performance of generative models like CycleGAN is known to be sensitive to hyperparameter tuning. The inconsistent performance of our hybrid TransConfMix approach suggests that naively combining pixel-level and feature-level methods is not a guaranteed path to improvement and requires more sophisticated integration.

As future work, we suggest (*i*) extend the evaluation to multiclass detection and report variability across multiple runs, (*ii*) analyze the impact of translation quality and resolution on detection performance, and (*iii*) explore a more integrated hybrid strategy by redesigning the ConfMix adaptation stage to jointly leverage three inputs — the source, the CycleGAN-translated source, and the target frames — within the same adaptation cycle. We also plan to investigate adaptations involving infrastructure-based sensors (e.g., roadside units) to further broaden the coverage of traffic scenes.

Acknowledgment

We gratefully acknowledge the support provided by the Brazilian agencies: Foundation of Research Support - Fundep (Conecta 2030, Rota 2030/Linha V, grant 29271.02.01/2022.04-00), São Paulo Research Foundation - FAPESP (grants 2023/17577-0 and 2024/22985-3), and National Council for Scientific and Technological Development - CNPq (grants 315220/2023-6 and 420442/2023-5).

References

[1] S. Cao, "Review of object detection challenges in autonomous driving," *Applied and Computational Engineering*, vol. 8, pp. 725–731, 08 2023.

[2] V. F. Arruda, R. F. Berriel, T. M. Paixão, C. Badue, A. F. De Souza, N. Sebe, and T. Oliveira-Santos, "Cross-domain object detection using unsupervised image translation," *Expert Systems with Applications*, vol. 192, p. 116334, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421016328

[3] F. Yu, D. Wang, Y. Chen, N. Karianakis, T. Shen, P. Yu, D. Lymberopoulos, S. Lu, W. Shi, and X. Chen, "SC-UDA: Style and content gaps aware unsupervised domain adaptation for object detection," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1061–1070.

[4] J. Wang, T. Shen, Y. Tian, Y. Wang, C. Gou, X. Wang, F. Yao, and C. Sun, "A parallel teacher for synthetic-to-real domain adaptation of traffic object detection," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 441–455, 2022.

[5] X. Wan, M. R. Sultan Mohd, J. Johari, and F. Ahmat Ruslan, "A review on object detection algorithms based deep learning methods," *Journal of Electrical & Electronic Systems Research*, pp. 1–13, 10 2023.

[6] R. M. Silva, G. F. Azevedo, M. V. Berto, J. R. Rocha, E. C. Fidelis, M. V. Nogueira, P. H. Lisboa, and T. A. Almeida, "Vulnerable road user detection and safety enhancement: A comprehensive survey," *Expert Systems with Applications*, vol. 292, p. 128529, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417425021487

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2020. [Online]. Available: https://arxiv.org/abs/1703.10593

[8] G. Mattolin, L. Zanella, E. Ricci, and Y. Wang, "ConfMix: Unsupervised domain adaptation for object detection via confidence-based mixing," 2022. [Online]. Available: https://arxiv.org/abs/2210.11539

[9] M. H. Amini and S. Nejati, "Bridging the gap between real-world and synthetic images for testing autonomous driving systems," 2024. [Online]. Available: https://arxiv.org/abs/2408.13950

[10] L. M. Kemeter, R. Hvingelby, P. Sierak, T. Schön, and B. Gosswam, "Towards reducing data acquisition and labeling for defect detection using simulated data," 2024. [Online]. Available: https://arxiv.org/abs/2406.19175

[11] X. Peng, B. Usman, K. Saito, N. Kaushik, J. Hoffman, and K. Saenko, "Syn2Real: A new benchmark forsynthetic-to-real visual domain adaptation," 2018. [Online]. Available: https://arxiv.org/abs/1806.09755

[12] J. Wang, T. Shen, Y. Tian, Y. Wang, C. Gou, X. Wang, F. Yao, and C. Sun, "A parallel teacher for synthetic-to-real domain adaptation of traffic object detection," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 441–455, 2022.

[13] R. Mao, J. Guo, Y. Jia, Y. Sun, S. Zhou, and Z. Niu, "DOLPHINS: Dataset for collaborative perception enabled harmonious and interconnected self-driving," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Cham: Springer Nature Switzerland, 2023, pp. 495–511.

[14] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?" in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 746–753.

[15] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Seattle, WA, USA: IEEE, 2020, pp. 11 621–11 631.