

FEDERAL UNIVERSITY OF SÃO CARLOS — UFSCAR
CENTER OF EXACT SCIENCES AND TECHNOLOGY — CCET
COMPUTING DEPARTMENT — DC
GRADUATE PROGRAM IN COMPUTER SCIENCE — PPGCC

André Neves Medeiros

**Simulation-to-Real Traffic Object
Detection: A Study of Pixel and
Feature-Level Unsupervised Domain
Adaptation**

André Neves Medeiros

**Simulation-to-Real Traffic Object
Detection: A Study of Pixel and
Feature-Level Unsupervised Domain
Adaptation**

Master's thesis presented to the Graduate Program in Computer Science at the Center of Exact Sciences and Technology of Federal University of São Carlos, as part of the requirements for obtaining the title of Master in Computer Science.

Concentration area: Computing Methodologies and Techniques

Advisor: Prof. Dr. Tiago A. Almeida

São Carlos

2026

This work is dedicated to my parents, for making possible what, for many, is not a choice but a privilege: the opportunity to study and to do science.

Acknowledgments

I would like to express my sincere gratitude to my advisor for the guidance, trust, and support throughout the development of this work.

I am also grateful to LaSID, the Intelligent Systems and Data Science Laboratory at UFSCar, for providing an academically stimulating environment that contributed to this research.

I am deeply grateful to my family for their constant support and for making my academic journey possible.

I thank my friends for their presence, encouragement, and for making this journey lighter.

I gratefully acknowledge the support provided by the Graduate Program in Computer Science at UFSCar (PPGCC/UFSCar), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, and the Brazilian agencies Foundation for 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).

“Sapere aude.”
(Horácio)

Resumo

O treinamento de modelos de detecção de objetos para aplicações no mundo real geralmente requer grandes volumes de dados anotados, cuja obtenção pode ser custosa e demorada. Conjuntos de dados sintéticos surgem como uma alternativa escalável, permitindo a geração automática de anotações e o controle sobre as condições de aquisição. No entanto, modelos treinados exclusivamente com dados sintéticos tendem a apresentar degradação de desempenho quando aplicados a imagens reais devido ao fenômeno conhecido como *domain shift*.

Esta dissertação investiga o problema de adaptação de domínio no cenário de transferência de simulação para o mundo real (*simulation-to-real*) aplicado à detecção de objetos, com foco na combinação de estratégias de adaptação em nível de pixel e em nível de características. No nível de pixel, o modelo CycleGAN é utilizado para traduzir imagens sintéticas para o estilo visual de conjuntos de dados reais, reduzindo discrepâncias de baixo nível entre domínios. No nível de características, o método ConfMix é adotado como abordagem base de adaptação de domínio não supervisionada, explorando a combinação entre amostras rotuladas do domínio fonte e pseudo-rótulos do domínio alvo.

Com base nessas abordagens, este trabalho propõe variantes denominadas TransConfMix e TransConfMix-2Stage que incorporam imagens sintéticas traduzidas ao processo de adaptação, além de uma extensão regularizada (ConfMix-Distill) que introduz um termo de divergência de Kullback–Leibler (KL) para incentivar a consistência entre previsões obtidas a partir de diferentes representações visuais de uma mesma amostra. Essas modificações têm como objetivo aumentar a estabilidade do treinamento e melhorar a generalização sob *domain shift*.

Os experimentos são conduzidos em múltiplos cenários de adaptação simulação-para-real, utilizando os conjuntos de dados DOLPHINS e SIM10K como domínios fonte, e Cityscapes e nuScenes como domínios alvo. Os resultados demonstram que a adaptação em nível de características melhora consistentemente o desempenho em relação ao treinamento apenas no domínio fonte nos cenários avaliados, enquanto a tradução em nível de

pixel pode atuar como suporte adicional quando a qualidade da tradução é suficientemente alta. No entanto, as extensões híbridas propostas não superaram de forma consistente o framework original ConfMix, indicando que a interação entre adaptação em nível de pixel e em nível de características é altamente sensível às características de cada domínio.

De forma geral, este trabalho apresenta uma análise sistemática da interação entre diferentes níveis de adaptação de domínio e mostra que os efeitos da combinação entre tradução em nível de pixel e adaptação em nível de características dependem fortemente da fidelidade da tradução e das características de cada domínio em tarefas de detecção de objetos.

Palavras-chave: Detecção de Objetos, Adaptação de Domínio, Adaptação de Domínio Não Supervisionada, Transferência Simulação-Real, Dados Sintéticos, CycleGAN, Adaptação em Nível de Características, Adaptação em Nível de Pixel, Destilação de Conhecimento.

Abstract

Training object detection models for real-world applications often requires large amounts of annotated data, which can be costly and time-consuming to obtain. Synthetic datasets provide a scalable alternative, enabling automatic annotation and controlled data generation. However, models trained exclusively on synthetic data typically suffer from performance degradation when applied to real-world images due to domain shift.

This dissertation investigates simulation-to-real domain adaptation for object detection, focusing on the combination of pixel-level and feature-level adaptation strategies. At the pixel level, CycleGAN is used to translate synthetic images into the visual style of real-world datasets, reducing low-level discrepancies between domains. At the feature level, the ConfMix framework is adopted as a baseline for unsupervised domain adaptation, leveraging confidence-based mixing between labeled source samples and pseudo-labeled target samples.

Building upon these approaches, this work proposes TransConfMix and TransConfMix-2Stage variants, which incorporate translated synthetic images into the adaptation process, and a KL-regularized extension (ConfMix-Distill) that enforces consistency between predictions obtained from original and translated versions of the same sample. These modifications aim to improve adaptation stability and enhance model generalization under domain shift.

Experiments are conducted on multiple synthetic-to-real scenarios, with DOLPHINS and SIM10K as source domains and Cityscapes and nuScenes as target domains. The results demonstrate that feature-level adaptation consistently improves over the source-only baseline across the evaluated scenarios, and that pixel-level translation can further support this process when translation quality is sufficiently high. However, the proposed hybrid extensions did not consistently surpass the original ConfMix framework, suggesting that the interaction between pixel- and feature-level adaptation is highly sensitive to domain characteristics and translation fidelity.

Overall, this work provides a systematic analysis of the interaction between pixel-

level and feature-level domain adaptation, showing that their combined effect in object detection is strongly dependent on translation fidelity and domain characteristics.

Keywords: Object Detection, Domain Adaptation, Unsupervised Domain Adaptation, Simulation-to-Real Transfer, Synthetic Data, CycleGAN, Feature-Level Adaptation, Pixel-Level Adaptation, Knowledge Distillation.

List of figures

Figure 1 – Examples of synthetic and real-world datasets used in this work. The top row shows synthetic images generated in simulated environments (Synthetic dataset for vehicle detection (SIM10K) and Dataset for Collaborative Perception enabled Harmonious and Interconnected Self-driving (DOLPHINS)), while the bottom row presents real-world images (Cityscapes dataset for urban scene understanding (Cityscapes) and nuScenes autonomous driving dataset (nuScenes)). Differences in texture realism, illumination, and scene complexity contribute to the Simulation-to-Real (Sim2Real) domain gap	40
Figure 2 – Overview of the ConfMix framework by Mattolin et al. (2023)	50
Figure 3 – Overview of the CycleGAN framework by Zhu et al. (2017). Two generators (G and F) learn mappings between domains X and Y , while discriminators (D_X and D_Y) enforce adversarial constraints. Cycle-consistency ensures that translated images can be mapped back to the original domain	54

Figure 4	– Overview of the proposed KL-regularized adaptation pipeline (ConfMix-Distill), highlighting the modifications introduced to the original ConfMix framework. The pipeline incorporates translated source images and an additional consistency constraint between predictions. The original synthetic image x_s and its translated counterpart x'_s (generated by Cycle-Consistent Adversarial Networks (CycleGAN)) are processed by two detectors: a teacher model $F^T(\theta)$, obtained after the first stage of ConfMix, and a student model $F(\theta)$ initialized from the teacher. The teacher produces predictions $\tilde{y}_s$ from x_s , while the student produces $\tilde{y}'_s$ from x'_s . A Kullback–Leibler (KL) divergence term L_{KL} is computed between these predictions to enforce consistency across visual domains. In parallel, the student is trained using the ground-truth annotations y_s through the detection loss L_{det} applied to $\tilde{y}'_s$. The adaptation process also includes the original ConfMix consistency term L_{cons} , computed using mixed samples. The final objective combines all loss components	62
Figure 5	– Example image from the DOLPHINS dataset used as a synthetic source domain	66
Figure 6	– Example image from the SIM10K dataset used as a synthetic source domain	67
Figure 7	– Example image from the Cityscapes dataset used as a real-world target domain	67
Figure 8	– Example image from the nuScenes dataset used as a real-world target domain	68
Figure 9	– Qualitative comparison of CycleGAN-based image translation. The two columns correspond to the DOLPHINS and SIM10K source datasets, while each row shows the original image and its translated versions in the Cityscapes and nuScenes domains	77
Figure 10	– Qualitative comparison of detection results for the SIM10K $\rightarrow$ Cityscapes scenario. Each column corresponds to a different adaptation method, showing representative predictions on the same target-domain image	80
Figure 11	– Qualitative comparison of detection results for the DOLPHINS $\rightarrow$ Cityscapes scenario. Each panel shows representative predictions obtained with the evaluated adaptation methods on the same target-domain image	81
Figure 12	– Qualitative comparison of detection results for the SIM10K $\rightarrow$ nuScenes scenario. Each panel shows representative predictions obtained with the evaluated adaptation methods on the same target-domain image . .	82

Figure 13 – Qualitative comparison of detection results for the DOLPHINS $\rightarrow$ nuScenes scenario. Each panel shows representative predictions obtained with the evaluated adaptation methods on the same target-domain image	84
Figure 14 – Example source images from the SIM10K and DOLPHINS datasets, shown for qualitative reference. These samples illustrate the type of input data used	106
Figure 15 – CycleGAN translation examples for the SIM10K $\rightarrow$ Cityscapes adaptation pair	107
Figure 16 – CycleGAN translation examples for the DOLPHINS $\rightarrow$ Cityscapes adaptation pair	108
Figure 17 – CycleGAN translation examples for the SIM10K $\rightarrow$ nuScenes adaptation pair	109
Figure 18 – CycleGAN translation examples for the DOLPHINS $\rightarrow$ nuScenes adaptation pair	110

List of tables

Table 1 – Comparison of the training protocols of ConfMix and the proposed variants. In all cases the second stage also uses the unlabeled target data $\mathcal{D}_T$. The last column indicates whether the KL-based consistency term is added.	63
Table 2 – Number of annotated vehicle instances in each dataset	68
Table 3 – Selected CycleGAN configurations for each domain adaptation scenario	76
Table 4 – Detection performance for the SIM10K $\rightarrow$ Cityscapes adaptation scenario	78
Table 5 – Detection performance for the DOLPHINS $\rightarrow$ Cityscapes adaptation scenario	80
Table 6 – Detection performance for the SIM10K $\rightarrow$ nuScenes adaptation scenario	81
Table 7 – Detection performance for the DOLPHINS $\rightarrow$ nuScenes adaptation scenario	83
Table 8 – Detection performance, additional metrics, and computational cost obtained using CycleGAN-translated SIM10K images evaluated on Cityscapes	107
Table 9 – Detection performance, additional metrics, and computational cost obtained using CycleGAN-translated DOLPHINS images evaluated on Cityscapes	108
Table 10 – Detection performance, additional metrics, and computational cost obtained using CycleGAN-translated SIM10K images evaluated on nuScenes	108
Table 11 – Detection performance, additional metrics, and computational cost obtained using CycleGAN-translated DOLPHINS images evaluated on nuScenes	109
Table 12 – Additional evaluation metrics for the SIM10K $\rightarrow$ Cityscapes scenario. Precision, recall, and F1-score are computed at IoU=0.5	113
Table 13 – Additional evaluation metrics for the DOLPHINS $\rightarrow$ Cityscapes scenario. Precision, recall, and F1-score are computed at IoU=0.5	113

Table 14 – Additional evaluation metrics for the SIM10K $\rightarrow$ nuScenes scenario.	
Precision, recall, and F1-score are computed at IoU=0.5	113
Table 15 – Additional evaluation metrics for the DOLPHINS $\rightarrow$ nuScenes scenario.	
Precision, recall, and F1-score are computed at IoU=0.5	114

List of symbols

δ	weighting factor controlling the transition between detector confidence and combined confidence
γ	dynamic weighting factor for the consistency loss term
γ_{id}	weighting factor controlling the importance of the identity loss
λ	weighting factor controlling the importance of the cycle consistency loss
λ_{KL}	weighting factor controlling the contribution of the KL divergence regularization term
$\mathcal{L}$	task loss function
$\mathcal{L}_{cyc}$	cycle consistency loss enforcing reconstruction of the original image after forward and backward translation
$\mathcal{L}_{GAN}$	adversarial loss used to align generated images with the target domain distribution
$\mathcal{L}_{id}$	identity loss encouraging generators to preserve input images when already in the target domain
Σ_i	predicted variance of the i -th bounding box parameter
C_{bbx}	bounding box confidence based on localization uncertainty
C_{comb}	combined confidence score used for pseudo-label selection
C_{det}	detector classification confidence score
D_X	discriminator that distinguishes real samples from X and generated samples $F(y)$
D_Y	discriminator that distinguishes real samples from Y and generated samples $G(x)$

D_{KL}	Kullback–Leibler divergence between two probability distributions
E	expectation operator
F	generator mapping images from domain Y to domain X
f_θ	model parameterized by θ
G	generator mapping images from domain X to domain Y
L_{cons}	consistency loss computed on mixed samples using pseudo-label supervision
L_{det}	supervised detection loss computed on labeled source data
L_{KL}	Kullback–Leibler divergence loss used for distillation between teacher and student predictions
L_{total}	overall training loss of the model
M_T	binary mask indicating the selected region from the target image
$p_S^{b,i,j,a,s}$	student objectness probability for batch element b , grid cell (i, j) , anchor a , and scale s
$p_T^{b,i,j,a,s}$	teacher objectness probability for batch element b , grid cell (i, j) , anchor a , and scale s
X	source image domain
x'_s	CycleGAN-translated version of the source image x_s
x_M	mixed image obtained by combining source and target regions
Y	target image domain

List of acronyms

CAD Computer-Aided Design

CARLA Car Learning to Act simulator

Cityscapes Cityscapes dataset for urban scene understanding

CNN Convolutional Neural Network

CycleGAN Cycle-Consistent Adversarial Networks

DA Domain Adaptation

DANN Domain-Adversarial Neural Network

DOLPHINS Dataset for Collaborative Perception enabled Harmonious and Interconnected Self-driving

FID Fréchet Inception Distance

GAN Generative Adversarial Network

IoU Intersection over Union

KITTI KITTI Vision Benchmark Suite

KL Kullback–Leibler

LiDAR Light Detection and Ranging

MMD Maximum Mean Discrepancy

MSCOCO Microsoft Common Objects in Context

mAP mean Average Precision

nuScenes nuScenes autonomous driving dataset

PASCALVOC PASCAL Visual Object Classes

RCNN Region-based Convolutional Neural Network

RPN Region Proposal Network

SIBGRAPI Conference on Graphics, Patterns and Images

SIM10K Synthetic dataset for vehicle detection

Sim2Real Simulation-to-Real

SSD Single Shot MultiBox Detector

Syn2Real Synthetic-to-Real benchmark

UDA Unsupervised Domain Adaptation

YOLO You Only Look Once

YOLOv5 You Only Look Once version 5

YOLOv8 You Only Look Once version 8

Contents

1	INTRODUCTION	29
1.1	Contextualization	29
1.2	Problem Statement	31
1.3	Research Gap	32
1.4	Hypothesis	33
1.5	Objectives	33
1.5.1	General Objective	34
1.6	Contributions	34
1.7	Research Questions	35
1.8	Dissertation Organization	36
2	THEORETICAL BACKGROUND	37
2.1	Object Detection	37
2.1.1	One-stage Detectors	38
2.2	Synthetic Data for Computer Vision	40
2.3	Domain Shift	41
2.4	Unsupervised Domain Adaptation	42
2.4.1	Pixel-Level Adaptation	44
2.4.2	Feature-Level Adaptation	45
3	RELATED WORK	47
3.1	Overview of Domain Adaptation Methods for Object Detection	47
3.2	ConfMix for Feature Level UDA	49
3.2.1	Method Overview	50
3.2.2	Confidence Estimation	50
3.2.3	Confidence-Based Region Mixing	51
3.2.4	Progressive Pseudo-Labeling	51

3.2.5	Training Objective	52
3.2.6	Results Reported by the Authors	52
3.2.7	Discussion	53
3.3	CycleGAN for Unpaired Image-to-Image Translation	53
3.3.1	Model Architecture	53
3.3.2	Training Objective	54
3.3.3	Applications and Limitations in Domain Adaptation	55
4	TRANSCONFMIX AND CONFMIX-DISTILL	57
4.1	TransConfMix	57
4.1.1	TransConfMix-2Stage	59
4.2	ConfMix-Distill	60
4.3	Final Remarks	63
5	METHODOLOGY	65
5.1	Datasets	65
5.1.1	Synthetic Datasets	66
5.1.2	Real-World Datasets	67
5.1.3	Summary of Datasets	68
5.2	Domain Adaptation Scenarios	68
5.3	Training Configuration	69
5.4	Reference Models	70
5.5	CycleGAN Hyperparameter Search	71
5.6	Compared Methods	71
5.7	Evaluation Metrics	72
6	RESULTS AND ANALYSIS	75
6.1	CycleGAN Hyperparameter Evaluation	75
6.2	Comparison of Domain Adaptation Methods	77
6.2.1	SIM10K -> Cityscapes	78
6.2.2	DOLPHINS -> Cityscapes	79
6.2.3	SIM10K -> nuScenes	81
6.2.4	DOLPHINS -> nuScenes	82
6.3	Summary of Results	84
7	CONCLUSION	87
7.1	Summary of Findings	87
7.2	Research Questions and Hypotheses Revisited	88
7.3	Limitations	92
7.4	Future Work	94

REFERENCES	97
APPENDIX A	DETAILED CYCLEGAN HYPERPARAMETER SEARCH
	105
A.1	Evaluated Configurations
	105
A.2	Complete Results by Domain Pair
	106
A.2.1	SIM10K -> Cityscapes
	106
A.2.2	DOLPHINS -> Cityscapes
	107
A.2.3	SIM10K -> nuScenes
	108
A.2.4	DOLPHINS -> nuScenes
	109
APPENDIX B	ADDITIONAL RESULTS AND IMPLEMENTATION DETAILS
	111
B.1	Hardware and Computational Infrastructure
	111
B.1.1	gpu1 (Primary Server)
	111
B.1.2	ws2 (Secondary Server)
	111
B.1.3	ss2 (Auxiliary Server)
	112
B.2	Additional Quantitative Results
	112
B.2.1	SIM10K to Cityscapes
	112
B.2.2	DOLPHINS to Cityscapes
	113
B.2.3	SIM10K to nuScenes
	113
B.2.4	DOLPHINS to nuScenes
	113

Chapter 1

Introduction

This section introduces the context of the problem addressed in this dissertation. It presents the main challenges associated with object detection under domain shift, motivates the use of synthetic data in computer vision, and outlines the need for effective domain adaptation strategies in simulation-to-real scenarios.

1.1 Contextualization

Object detection has experienced substantial progress over the last decade, largely driven by advances in deep convolutional neural networks and the availability of large annotated datasets. Modern detection architectures, such as Faster Region-based Convolutional Neural Network (RCNN) (REN et al., 2015), You Only Look Once (YOLO) (REDMON et al., 2016), and subsequent variants, have achieved state-of-the-art performance on benchmark datasets including PASCAL Visual Object Classes (PASCALVOC) and Microsoft Common Objects in Context (MSCOCO). These models typically assume that training and testing samples are drawn from similar data distributions. However, when deployed in real-world environments, object detectors frequently suffer from performance degradation due to distributional discrepancies between the training data and the operational domain, a phenomenon commonly known as domain shift (GANIN et al., 2016).

In many real-world computer vision applications, the collection and annotation of large-scale datasets can be expensive, labor-intensive, and in some cases impractical. To mitigate these limitations, synthetic data generation has become an increasingly adopted strategy for training machine learning models. Modern simulation environments and rendering pipelines allow the generation of large datasets with precise ground-truth an-

notations at virtually no manual labeling cost. Such environments enable fine-grained control over scene parameters, including object placement, illumination conditions, camera viewpoints, and background composition. Consequently, synthetic data has proven particularly valuable in scenarios involving safety-critical systems, industrial inspection, and infrastructure monitoring, where real-world data acquisition may be constrained by operational risks, privacy concerns, or regulatory restrictions (TOBIN et al., 2017; PENG et al., 2018b; TREMBLAY et al., 2018).

However, despite these advantages, models trained on synthetic data often fail to generalize effectively to real-world environments due to distributional discrepancies between the simulated and real domains, a phenomenon commonly referred to as *domain shift*. These discrepancies arise from differences in low-level visual properties, such as texture statistics, illumination conditions, and sensor noise, as well as higher-level factors, including object appearance and scene complexity. As a result, the learned feature representations may not transfer reliably across domains, leading to significant performance degradation when models are deployed in real-world scenarios (TORRALBA; EFROS, 2011; TZENG et al., 2017).

To address this limitation, Domain Adaptation (DA) techniques aim to reduce the discrepancy between source and target distributions by learning domain-invariant representations (WANG; DENG, 2018). In computer vision tasks, a widely studied setting is Unsupervised Domain Adaptation (UDA), where labeled data is available only in the source domain while the target domain remains unlabeled. Under this setting, adaptation strategies seek to transfer knowledge from the labeled source domain to the unlabeled target domain without requiring additional manual annotations (GANIN et al., 2016).

Another complementary strategy for addressing domain discrepancies relies on pixel-level domain translation. Generative models based on adversarial learning, particularly CycleGAN, enable unpaired image-to-image translation between domains by jointly enforcing adversarial alignment and cycle-consistency constraints (ZHU et al., 2017). By transforming synthetic images to resemble the visual characteristics of real-world data, these approaches aim to reduce low-level discrepancies such as texture statistics, illumination patterns, and color distributions. Pixel-level translation methods are often used either as a preprocessing stage or in combination with feature-level adaptation strategies to improve cross-domain generalization (HOFFMAN et al., 2018).

Building upon the challenges associated with domain shift between synthetic and real data, the Sim2Real transfer problem has emerged as a central research topic in computer vision and robotics. In this context, Sim2Real transfer explicitly addresses the difficulty of deploying models trained in simulated environments into real-world scenarios. Although synthetic environments enable scalable data generation and precise annotations, models trained exclusively in simulation frequently struggle to generalize to real-world observations due to the persistent gap between simulated and physical environments (TOBIN et

al., 2017; PENG et al., 2018b). Bridging this gap requires adaptation mechanisms capable of reducing distributional divergence while preserving the semantic structure necessary for reliable visual recognition. Addressing this challenge motivates the investigation conducted in this work.

This dissertation is developed within the context of the Conecta 2030 project, a research initiative aimed at improving road safety through the automatic perception of traffic scenes. In such applications, the reliable detection of road users is essential to support safety-critical decisions, for instance at pedestrian crossings, where perception failures may have severe consequences. Acquiring and annotating large volumes of real-world traffic data for this purpose is costly and operationally constrained, which makes synthetic data and Sim2Real transfer particularly attractive in this domain. Within this broader effort, the present work focuses on vehicle detection as a controlled and well-defined case, since the vehicle class is consistently annotated across the synthetic and real datasets considered; nevertheless, the adaptation strategies investigated here are general and can be extended to other categories of road users. Addressing the Sim2Real domain gap in this setting is therefore the central challenge that motivates the investigation conducted in this work.

1.2 Problem Statement

This dissertation addresses the problem of adapting object detection models trained on synthetic data to real-world domains. In the investigated setting, labeled images are available in the source domain, while the target domain is composed of unlabeled real-world images. Although both domains contain the same semantic categories of interest, they differ substantially in visual appearance, scene composition, and imaging conditions. As a result, detectors trained only on synthetic data often suffer a significant drop in performance when applied to real-world images.

In practical terms, this degradation may appear as lower detection accuracy, reduced recall, and less reliable confidence estimates. The problem becomes particularly relevant in simulation-to-real scenarios, where synthetic data offers scalable and inexpensive supervision, but real-world deployment requires robustness to variations that are not fully represented in simulated environments.

From the perspective of domain adaptation, the central challenge is to reduce the discrepancy between source and target domains without relying on labeled target data. In the UDA setting adopted in this work, adaptation strategies seek to exploit labeled synthetic samples together with unlabeled real-world images in order to learn representations that generalize more effectively across domains. A more formal treatment of this setting is presented in Chapter 2.

Despite recent progress in both pixel-level and feature-level adaptation methods, inte-

grating these strategies into modern one-stage object detectors remains challenging. Architectural design choices, training stability, and the need to preserve detection-specific properties can complicate the adaptation process.

Therefore, the problem addressed in this dissertation can be summarized as follows:

How can object detection models trained on synthetic data be effectively adapted to real-world domains under limited or absent target supervision, while preserving detection performance and training stability?

1.3 Research Gap

Although significant progress has been made in DA for computer vision, important limitations remain when addressing Sim2Real transfer in object detection.

Pixel-level adaptation methods, such as CycleGAN-based image translation (ZHU et al., 2017), attempt to reduce low-level discrepancies between domains by transforming synthetic images to resemble real-world samples. While these approaches can reduce differences in texture and illumination, they also introduce practical challenges. First, image translation models are often sensitive to hyperparameter choices and architectural configurations, which may result in unstable training or visual artifacts that compromise semantic consistency. Second, improvements in visual realism do not necessarily translate into better object detection performance, particularly when structural relationships between objects and their surrounding context are not fully preserved during translation.

On the other hand, feature-level adaptation strategies aim to align intermediate representations within the detection network. Methods based on consistency regularization and confidence-guided pseudo-labeling, such as ConfMix (MATTO LIN et al., 2023), attempt to exploit unlabeled target data by encouraging prediction stability under data mixing or perturbation strategies. While promising, these approaches depend strongly on the reliability of pseudo-labels and on stable confidence estimation. When pseudo-label quality deteriorates, errors may propagate through the adaptation process, which can be particularly problematic for complex one-stage detectors.

Furthermore, most previous studies evaluate pixel-level and feature-level adaptation separately, without systematically analyzing how the quality of image translation influences the stability of the adaptation process. In particular, limited attention has been given to how generative model calibration affects downstream adaptation, or to how additional regularization may improve the robustness of confidence-driven adaptation frameworks.

In the context of modern object detectors such as You Only Look Once version 8 (YOLOv8), integrating domain translation and adaptation mechanisms introduces additional architectural challenges. Detection heads, anchor-free designs, and multi-scale prediction layers create dependencies that may amplify the effects of domain misalignment.

As a result, adaptation techniques originally developed for classification or segmentation tasks cannot always be directly applied to object detection without modifications.

Therefore, a clear research gap emerges: there is a need for a systematic investigation of (i) how high-quality pixel-level domain translation impacts Sim2Real object detection, and (ii) how consistency-based adaptation frameworks can be enhanced to improve stability and generalization under distributional shift. Addressing this gap requires not only empirical evaluation but also architectural refinement and loss-level regularization tailored to detection-specific dynamics.

1.4 Hypothesis

This work is guided by the following central hypothesis:

The performance degradation observed when transferring object detection models from simulation to real-world environments can be significantly reduced by combining high-quality pixel-level domain translation with improved consistency-based feature adaptation.

More specifically, the following sub-hypotheses are investigated:

- **H1:** Systematic optimization of CycleGAN hyperparameters and architectural configurations improves the semantic consistency between translated synthetic images and real target samples, which in turn benefits downstream object detection performance.
- **H2:** Incorporating an additional Kullback–Leibler (KL) divergence regularization term into a consistency-based adaptation framework improves prediction calibration and reduces the propagation of errors during pseudo-label refinement.
- **H3:** Pixel-level alignment and feature-level consistency regularization provide complementary advantages, and their joint application leads to better Sim2Real transfer performance than either strategy used independently.

These hypotheses are evaluated through controlled experiments in a Sim2Real object detection setting. The experiments use You Only Look Once version 5 (YOLOv5) as the base detection architecture and compare baseline training, translation-based adaptation, and extended consistency-based adaptation strategies.

1.5 Objectives

This section presents the objectives of this research, outlining both the general goal and the specific objectives that guide the development of this work.

1.5.1 General Objective

The main objective of this dissertation is to investigate methods for reducing the Sim2Real domain gap in object detection by integrating pixel-level domain translation and confidence-guided feature-level adaptation within a modern detection framework.

The specific objectives are presented as follows:

- ❑ Assess the impact of Sim2Real domain shift on object detection performance when models are trained exclusively on synthetic data and applied to real-world datasets.
- ❑ Study pixel-level DA using CycleGAN, examining how image translation quality influences downstream detection performance.
- ❑ Perform a systematic exploration of CycleGAN hyperparameters and generator configurations to improve the quality of synthetic-to-real image translation.
- ❑ Implement and evaluate a feature-level adaptation strategy based on confidence-guided data mixing following the ConfMix framework.
- ❑ Extend the adaptation framework by introducing a Kullback–Leibler (KL) divergence regularization term aimed at improving prediction stability and confidence calibration.
- ❑ Conduct an experimental comparison between baseline training, pixel-level adaptation, and enhanced feature-level adaptation strategies in a Sim2Real setting.
- ❑ Analyze the effectiveness, limitations, and interactions between pixel-level translation and feature-level adaptation mechanisms for object detection under domain shift.

1.6 Contributions

This dissertation presents a comprehensive investigation of Sim2Real DA for object detection. The results presented here are derived from the research conducted throughout this master’s program, including intermediate findings previously published in Conference on Graphics, Patterns and Images (SIBGRAPI) by Medeiros et al. (2025).

In comparison to earlier results, this dissertation provides a more extensive experimental evaluation, explores additional DA scenarios, and introduces methodological refinements aimed at improving both adaptation stability and translation quality.

The main contributions of this work are summarized as follows:

- ❑ **Systematic evaluation of sim-to-real DA strategies for object detection.**
A comprehensive empirical analysis of when and how pixel-level domain translation

interacts with confidence-guided feature-level adaptation. The study demonstrates that this interaction is strongly conditioned on translation quality, source domain characteristics, and domain pair complexity — findings that provide actionable guidelines for practitioners and motivate future research on jointly optimizing translation and adaptation.

- ❑ **Optimization of CycleGAN-based domain translation.** A systematic exploration of CycleGAN training configurations is performed to improve the quality of Sim2Real image translation. The analysis investigates how generator architecture and training parameters affect the semantic consistency of translated images and their influence on downstream detection performance.
- ❑ **Extension of the ConfMix adaptation framework with KL-divergence regularization.** An extension to the confidence-based mixing strategy is proposed through the introduction of a Kullback–Leibler (KL) divergence regularization term. This modification aims to improve prediction stability and reduce the propagation of noisy pseudo-labels during adaptation.
- ❑ **Comprehensive analysis of adaptation dynamics in modern object detectors.** The study examines how pixel-level translation quality and feature-level adaptation interact when applied to contemporary object detection architectures, providing insights into their limitations and complementary effects in Sim2Real transfer scenarios.

1.7 Research Questions

Based on the research gap identified in the previous sections, this dissertation investigates the following research questions:

- ❑ **RQ1:** To what extent does the Sim2Real domain shift affect the performance of object detectors trained exclusively on synthetic datasets?
- ❑ **RQ2:** How does the quality of pixel-level domain translation influence object detection performance in Sim2Real transfer scenarios?
- ❑ **RQ3:** Can consistency-based DA frameworks be improved by introducing additional regularization mechanisms, such as KL-divergence constraints, to stabilize pseudo-label learning?
- ❑ **RQ4:** How do pixel-level domain translation and feature-level adaptation strategies interact when applied to modern object detection architectures in Sim2Real settings?

These research questions guide the experimental design of this study and serve as the basis for evaluating the effectiveness of the proposed methodological refinements in reducing the Sim2Real domain gap.

1.8 Dissertation Organization

The remainder of this dissertation is organized as follows.

Chapter 2 presents the theoretical background necessary to support this study. It reviews object detection methods, DA techniques, and previous approaches related to Sim2Real transfer in computer vision.

Chapter 3 presents the related work most closely connected to the problem investigated in this dissertation, with emphasis on pixel-level and feature-level adaptation strategies for object detection under domain shift.

Chapter 4 presents the methods proposed in this dissertation, namely TransConfMix and ConfMix-Distill, together with their respective variants and formulations.

Chapter 5 describes the experimental methodology adopted in this work. It details the datasets, domain adaptation scenarios, training configuration, evaluated methods, and model selection procedures used in the experimental pipeline.

Chapter 6 presents and analyzes the experimental results. Both quantitative and qualitative evaluations are used to assess the effectiveness of the investigated adaptation strategies and the impact of the proposed methods.

Finally, Chapter 7 concludes the dissertation by summarizing the main findings, discussing the limitations of the study, and outlining directions for future research.

Chapter 2

Theoretical Background

This section presents the fundamental concepts of object detection in computer vision. It introduces the main paradigms and architectures used for detecting objects in images, providing the necessary background to understand the methods and adaptation strategies explored in this dissertation.

2.1 Object Detection

Object detection is a fundamental task in computer vision that focuses on identifying and locating objects of interest within images or videos. In practical terms, this means not only recognizing what objects are present, but also determining where they appear in the scene.

This task differs from image classification, where a single label is assigned to the entire image. In object detection, multiple objects can be present simultaneously, each belonging to potentially different categories and occupying distinct spatial regions. To represent this information, object detectors typically use bounding boxes, which define the position and extent of each detected object in the image.

The ability to detect and localize objects is essential in a wide range of real-world applications. Examples include autonomous driving, where vehicles and pedestrians must be identified in real time; robotics, where systems need to interact with their environment; and surveillance and transportation systems, where monitoring and analysis of dynamic scenes are required (SULTANA; SUFIAN; DUTTA, 2019; ZOU et al., 2023).

Over the past decade, object detection has undergone significant advances, largely driven by the development of deep learning techniques, particularly Convolutional Neural Networks (CNNs). These models are designed to automatically learn visual features from

data, eliminating the need for manually engineered descriptors that were common in earlier approaches (LECUN; BENGIO; HINTON, 2015).

A key characteristic of CNNs is their ability to learn hierarchical representations. In this process, lower layers of the network capture simple visual patterns such as edges and textures, while deeper layers encode more complex structures, including object parts and semantic concepts. This hierarchical learning enables models to effectively recognize and localize objects under a wide range of visual conditions.

As a result of these advances, modern object detection systems have achieved substantial improvements in both accuracy and computational efficiency. Consequently, object detection has become a central component of contemporary visual perception systems, being widely adopted in real-time and large-scale applications (SAPKOTA; KARKEE, 2025; REN et al., 2015; REDMON et al., 2016).

Deep learning-based object detectors are categorized into two main paradigms: two-stage detectors and one-stage detectors. This distinction is based on how the detection process is structured within the model (LIU et al., 2020).

Two-stage detectors divide the task into two sequential steps. First, the model identifies regions in the image that are likely to contain objects, known as region proposals. Then, in a second stage, these candidate regions are classified and refined to produce the final detections.

A representative example of this paradigm is the RCNN family. In the original RCNN framework, region proposals are generated using an external algorithm called selective search. These regions are then individually processed by a convolutional neural network to extract features and perform classification (GIRSHICK et al., 2014). While effective, this approach is computationally costly due to the repeated processing of multiple regions.

To address this limitation, subsequent methods introduced more efficient architectures. In particular, Faster RCNN incorporates the Region Proposal Network (RPN), which generates region proposals directly within the neural network. This integration allows both proposal generation and object detection to be learned jointly, significantly improving efficiency while maintaining strong detection performance (REN et al., 2015; SULTANA; SUFIAN; DUTTA, 2019).

2.1.1 One-stage Detectors

In contrast to two-stage approaches, one-stage detectors perform object localization and classification in a single forward pass of the network. This means that the model directly predicts the positions and categories of objects without relying on an intermediate step to generate candidate regions.

By removing the explicit region proposal stage, one-stage detectors significantly simplify the detection pipeline. Instead of processing multiple candidate regions separately, the model analyzes the entire image at once and produces all predictions simultaneously.

Representative models in this category include YOLO and Single Shot MultiBox Detector (SSD). These architectures are designed to achieve high computational efficiency, making them suitable for real-time applications. Although earlier one-stage detectors exhibited lower accuracy compared to two-stage methods, subsequent improvements have enabled them to reach competitive performance while maintaining substantially faster inference speeds (REDMON et al., 2016; LIU et al., 2016; REDMON; FARHADI, 2018).

YOLO formulates object detection as a unified regression problem. Instead of processing candidate regions separately, the input image is divided into a grid, and the model directly predicts bounding boxes and class probabilities for each region in a single step. This design enables high computational efficiency, making YOLO suitable for real-time applications while still achieving strong detection accuracy (REDMON et al., 2016).

Similarly, SSD performs detection at multiple scales by leveraging feature maps extracted from different layers of the network. Since shallow layers capture fine spatial details and deeper layers encode higher-level semantic information, combining these representations allows the model to detect objects of varying sizes more effectively. This multi-scale design provides a good balance between speed and precision (LIU et al., 2016; SULTANA; SUFIAN; DUTTA, 2019).

Due to their efficiency and scalability, one-stage detectors have been widely adopted in real-world applications. Over time, the YOLO family has evolved through several architectural improvements, progressively enhancing detection accuracy, inference speed, and deployment efficiency. As a result, YOLO-based models have become one of the most influential frameworks for real-time object detection in modern computer vision systems (SAPKOTA; KARKEE, 2025; WANG; BOCHKOVSKIY; LIAO, 2023).

Despite these advances in detection architectures, training robust object detectors still requires large amounts of annotated data. In detection tasks, this process is particularly demanding, as it involves not only identifying object categories but also precisely defining their spatial locations through bounding boxes.

The creation of large-scale labeled datasets is therefore both costly and time-consuming. For this reason, several studies have explored the use of synthetic datasets as a scalable alternative for training deep learning models. Synthetic data allows automatic generation of annotations, significantly reducing manual effort.

However, models trained exclusively on synthetic data often experience performance degradation when applied to real-world images. This limitation arises from differences between simulated and real environments, commonly referred to as the Sim2Real domain gap, as illustrated in Figure 1 (PENG et al., 2018b).



SIM10K (synthetic)



DOLPHINS (synthetic)



Cityscapes (real)



nuScenes (real)

Figure 1 – Examples of synthetic and real-world datasets used in this work. The top row shows synthetic images generated in simulated environments (SIM10K and DOLPHINS), while the bottom row presents real-world images (Cityscapes and nuScenes). Differences in texture realism, illumination, and scene complexity contribute to the Sim2Real domain gap

2.2 Synthetic Data for Computer Vision

Deep learning models for computer vision typically require large amounts of annotated data to achieve high performance. In tasks such as object detection, this requirement becomes even more demanding, as it involves not only identifying object categories but also precisely defining their spatial locations within the image. As a result, the creation of high-quality labeled datasets is often expensive and time-consuming, since it relies on manual annotation across large collections of images. This dependency has motivated the exploration of alternative approaches for generating training data, including the use of synthetic data produced in simulated environments (VAROL et al., 2017).

One promising strategy is the use of synthetic data, which can be generated using computer graphics engines, simulation platforms, or procedural generation techniques. These approaches allow images to be created together with precise annotations, since the entire scene generation process is controlled. In this context, labels such as bounding boxes, segmentation masks, depth maps, and optical flow can be automatically obtained without manual intervention. Several synthetic datasets have been proposed to support research in computer vision, including large-scale collections generated from virtual environments and computer games (RICHTER et al., 2016; ROS et al., 2016).

The use of simulation environments offers several advantages for training computer vision models. Synthetic scenes allow precise control over environmental factors such as illumination conditions, object placement, camera viewpoint, and overall scene complexity. This level of control makes it possible to systematically generate diverse training samples, which can improve model robustness. In addition, it avoids many of the logistical, financial, and safety challenges associated with collecting real-world data. As a result, synthetic data has been widely explored in applications such as autonomous driving, robotics, and industrial inspection (VAROL et al., 2017).

Despite these advantages, models trained exclusively on synthetic data often experience a noticeable performance drop when applied to real-world images. This occurs because simulated environments rarely reproduce all visual characteristics of real-world scenes with complete fidelity. Differences in texture realism, lighting conditions, sensor noise, and scene variability introduce discrepancies between synthetic and real image distributions, making it difficult for models to generalize across domains.

This effect is often described as the sim-to-real gap or the reality gap. To mitigate this issue, several strategies have been proposed in the literature. One approach is domain randomization, which exposes models to a wide range of visual variations during training—such as changes in textures, lighting, and object appearances—so that real-world observations can be interpreted as another variation of the simulated environment (TOBIN et al., 2017; TREMBLAY et al., 2018). Another line of research focuses on DA techniques that explicitly align the distributions of synthetic and real data, enabling models trained in simulation to generalize more effectively to real-world environments.

2.3 Domain Shift

Although modern deep learning models have achieved remarkable performance in computer vision tasks, their success often relies on the assumption that training and testing data are drawn from the same underlying distribution. In simpler terms, this means that the data used to train the model should be similar to the data it will encounter during deployment. In practice, however, this assumption is frequently violated. When models trained on one dataset are applied to different environments or evaluated on new datasets, their performance may degrade due to differences in visual characteristics. This phenomenon is referred to as domain shift.

Domain shift has been extensively studied in the computer vision community. One well-known manifestation of this problem is dataset bias, where models learn patterns that are specific to the training dataset rather than generalizable visual representations. As a result, a model may perform well on the dataset it was trained on, but fail when evaluated on another dataset with different object appearances, scene structures, or imaging conditions (TORRALBA; EFROS, 2011).

From a theoretical perspective, domain shift occurs when the joint probability distribution of inputs and labels differs between the source and target domains. In other words, the relationship between images and their corresponding labels changes across domains. Formally, if $P_S(X, Y)$ denotes the joint distribution of the source domain and $P_T(X, Y)$ denotes the distribution of the target domain, domain shift arises when $P_S(X, Y) \neq P_T(X, Y)$. Under these conditions, minimizing the empirical risk on the source domain does not necessarily guarantee low risk in the target domain, meaning that a model trained on source data may fail to generalize to new environments (BEN-DAVID et al., 2010).

In computer vision applications, domain shift can arise from several factors, including variations in illumination, sensor characteristics, object appearance, background clutter, and environmental context. These differences affect how objects are represented in the image, even when the underlying semantic content remains the same. This effect becomes even more pronounced when models trained on synthetic data are deployed in real-world scenarios, where visual conditions often differ substantially from those observed in simulated environments.

To address these challenges, a large body of research has focused on methods that reduce the discrepancy between domains. These approaches belong to the broader field of domain adaptation, which aims to transfer knowledge learned from a labeled source domain to a target domain with limited or no labeled data (CSURKA, 2017; WANG; DENG, 2018). In general, DA methods seek to align feature distributions, reduce representation differences, or adapt model parameters so that knowledge learned from the source domain can generalize more effectively to the target domain.

2.4 Unsupervised Domain Adaptation

DA addresses the problem of performance degradation that arises when a model trained on a source domain is applied to a target domain with a different data distribution. In practical terms, this means that a model trained in one environment (e.g., synthetic images) may not perform well when deployed in another (e.g., real-world images). Formally, let $\mathcal{D}_S = \{(x_s, y_s)\}$ denote the labeled source domain and $\mathcal{D}_T = \{x_t\}$ the unlabeled target domain. In the UDA setting, the goal is to learn a model that generalizes well to $\mathcal{D}_T$, despite the discrepancy between the joint distributions $P_S(X, Y)$ and $P_T(X, Y)$.

In a supervised learning setting, let f_θ denote a model trained on the source domain. Training seeks to minimize the expected source risk

$$\mathbb{E}_{(X,Y) \sim P_S}[\mathcal{L}(f_\theta(X), Y)],$$

whereas the goal of adaptation is to achieve low risk on the target domain,

$$\mathbb{E}_{(X,Y) \sim P_T}[\mathcal{L}(f_\theta(X), Y)].$$

When $P_S(X, Y) \neq P_T(X, Y)$, minimizing the source risk does not necessarily guarantee low target risk, which is the central difficulty addressed by UDA.

Theoretical analyses have shown that the error on the target domain depends on three main factors: the error on the source domain, the divergence between source and target distributions, and the capacity of the model to adapt across domains (BEN-DAVID et al., 2010). This result highlights that achieving good performance in UDA requires not only learning from labeled source data, but also reducing the discrepancy between domains. In practice, such discrepancies may arise from differences in low-level properties, such as texture and illumination, as well as higher-level factors, including object appearance and scene composition. Consequently, UDA methods aim to bridge this gap so that knowledge learned from the source domain can be effectively transferred to the target domain.

Deep learning-based approaches have become the dominant paradigm for UDA in computer vision, as they enable models to learn representations that are both discriminative for the task and robust to domain-specific variations (LONG et al., 2015; GANIN; LEMPITSKY, 2015). In other words, the goal is to extract features that are useful for solving the task (e.g., detecting objects) while being less sensitive to differences between domains. A common taxonomy categorizes UDA methods into three main groups: discrepancy-based methods, adversarial methods, and self-training approaches.

Discrepancy-based methods aim to explicitly reduce the difference between feature distributions of the source and target domains. This is typically done by minimizing a statistical distance measure, such as Maximum Mean Discrepancy (MMD) or correlation alignment, between features extracted from both domains (LONG et al., 2015; SUN; SAENKO, 2016). By encouraging these distributions to become similar, the model learns representations that are less dependent on domain-specific characteristics and more transferable across domains.

Adversarial domain adaptation methods approach this problem from a different perspective, formulating domain alignment as a minimax optimization task. In this setting, a feature extractor is trained together with a domain discriminator: while the discriminator attempts to distinguish whether a feature comes from the source or target domain, the feature extractor is optimized to make this distinction difficult. This adversarial interaction encourages the learned features to become domain-invariant. A prominent example is the Domain-Adversarial Neural Network (DANN), which introduced the gradient reversal layer to enable joint optimization of feature learning and domain alignment (GANIN et al., 2016). Subsequent works have extended this idea by proposing alternative adversarial objectives and strategies for aligning decision boundaries between domains (TZENG et al., 2017; Saito, Kuniaki et al., 2018; Long, Mingsheng et al., 2018).

Another important class of approaches relies on self-training with pseudo-labels, where high-confidence predictions on target data are iteratively used as supervision signals (ZOU et al., 2018; XIE et al., 2020). In this setting, the model first generates predictions for

unlabeled target samples and then uses the most reliable ones as if they were ground-truth labels. Over successive training iterations, this process allows the model to gradually adapt to the target domain. These methods have shown strong performance, particularly when combined with confidence filtering mechanisms to discard uncertain predictions and with consistency regularization strategies that stabilize training.

In computer vision tasks such as object detection and semantic segmentation, DA presents additional challenges due to the structured nature of the outputs. Unlike classification, where a single label is predicted per image, detection models must simultaneously localize and classify multiple objects, each with its own position and category. This makes domain alignment more complex, as both spatial and semantic consistency must be preserved across domains.

To address these challenges, UDA methods for detection typically operate at multiple levels. Pixel-level methods aim to reduce low-level discrepancies by translating source images to resemble the target domain, often using generative models such as CycleGAN (ZHU et al., 2017). Feature-level approaches, on the other hand, focus on aligning intermediate representations within the network, for example through adversarial learning applied at different feature stages (CHEN et al., 2018; SAITO et al., 2019). Finally, output-level methods enforce consistency directly in the prediction space, encouraging similar structural outputs across domains (TSAI et al., 2018).

Each of these strategies presents different trade-offs. Pixel-level adaptation can effectively reduce visible differences between domains by making synthetic images resemble real ones; however, the translation process may introduce visual artifacts or distort important details, which can negatively affect downstream tasks. Feature-level methods, in contrast, operate on internal representations and therefore tend to provide more stable training, but they may struggle when the visual gap between domains is large and difficult to bridge through feature alignment alone. For this reason, hybrid approaches that combine multiple levels of adaptation have become increasingly common, as they seek to leverage the complementary strengths of different strategies and achieve more robust cross-domain generalization.

2.4.1 Pixel-Level Adaptation

Pixel-level DA seeks to reduce differences between domains directly in the image space. Instead of modifying internal representations learned by the model, these methods transform the input images themselves so that samples from the source domain visually resemble those from the target domain. The key idea is to change low-level properties, such as color, texture, and illumination, while preserving the semantic content of the scene. By doing so, the translated images can be used to train models on data that is visually closer to the target distribution.

Generative Adversarial Networks (GANs) have become a common choice for this type of adaptation. These models learn to generate images that follow the distribution of a target domain by training two components in opposition: a generator, which produces transformed samples, and a discriminator, which attempts to distinguish between source and target domains. In the context of DA, image-to-image translation models use this framework to map images from one domain to another (ISOLA et al., 2017; ZHU et al., 2017).

One of the main advantages of unpaired image translation methods is that they do not require corresponding images from both domains. This property makes them particularly suitable for Sim2Real scenarios, where paired data is rarely available. In computer vision, such methods have been widely used to translate synthetic images into more realistic ones, helping reduce the discrepancy between simulated and real-world data.

Despite these advantages, pixel-level adaptation presents important challenges. The image translation process may introduce visual artifacts or alter relevant details if not properly constrained. Moreover, improvements in visual realism do not necessarily translate into better performance in downstream tasks such as object detection. This occurs because models are sensitive not only to appearance, but also to the underlying structure of the data. For this reason, pixel-level adaptation is often combined with feature-level or output-level strategies to achieve more robust cross-domain generalization.

2.4.2 Feature-Level Adaptation

Feature-level DA seeks to reduce differences between domains by aligning the internal representations learned by deep neural networks. These representations, often referred to as features, correspond to intermediate activations of the network that encode visual patterns such as edges, textures, object parts, and higher-level semantic concepts. Instead of modifying the input images, these methods aim to learn feature embeddings that are less sensitive to domain-specific variations, allowing a model trained on source data to generalize more effectively to the target domain (WANG; DENG, 2018; GANIN et al., 2016; SUN; SAENKO, 2016).

Several strategies have been proposed to achieve feature-level alignment. Adversarial methods aim to learn representations that are indistinguishable across domains by training a feature extractor to confuse a domain discriminator. Discrepancy-based methods, on the other hand, explicitly minimize distance measures between feature distributions, encouraging statistical similarity between source and target representations. Other approaches rely on self-training, in which pseudo-labels generated for target-domain samples are iteratively refined and used as supervision signals during training (WANG; DENG, 2018).

In object detection, feature-level adaptation presents additional challenges due to the structured nature of the task. Detection models must simultaneously localize and classify

multiple objects, which requires preserving both spatial information and semantic consistency. As a result, DA methods must consider not only global feature alignment, but also the consistency of object-level predictions across domains (CHEN et al., 2018; SAITO et al., 2019).

Recent work has also explored confidence-based adaptation strategies that leverage pseudo-labels obtained from the target domain to guide the training process. In such approaches, only predictions that satisfy confidence criteria are incorporated into training, reducing the risk of introducing incorrect supervision.

Feature-level adaptation methods based on pseudo-labeling and consistency regularization have shown promising results in improving cross-domain object detection performance, particularly in Sim2Real scenarios. By encouraging consistency between predictions across domains and leveraging confidence-aware learning mechanisms, these approaches enable more stable adaptation when labeled target data is unavailable.

Chapter 3

Related Work

Research on DA for object detection has advanced significantly in recent years, particularly in the context of Sim2Real transfer. A wide range of approaches has been proposed to address the performance degradation caused by domain shift, exploring different strategies to align source and target domains at multiple levels.

3.1 Overview of Domain Adaptation Methods for Object Detection

Early research in DA primarily focused on feature-level alignment, aiming to learn domain-invariant representations by reducing discrepancies between source and target feature distributions. These approaches were later extended to object detection by incorporating alignment mechanisms into different stages of the detection pipeline, including image-level, instance-level, and multi-level features (CHEN et al., 2018; SAITO et al., 2019).

Subsequent studies have explored adaptations tailored to different detector architectures and application scenarios. For instance, Yu e Lu (2024) proposed a DA method for anchor-free object detection in urban traffic scenes, demonstrating that cross-domain adaptation can be effectively applied beyond traditional anchor-based detectors (YU; LU, 2024). In a complementary direction, other works have investigated the integration of synthetic dataset generation with adaptation strategies to improve detection performance in Sim2Real scenarios (SADEGHIGOL; KEYVANRAD, 2024).

From a methodological perspective, existing approaches can be broadly categorized into pixel-level, feature-level, and hybrid strategies. Pixel-level methods reduce low-level

discrepancies by transforming source images toward the target domain, while feature-level methods focus on aligning internal representations learned by the detector. Hybrid approaches combine both strategies in order to leverage their complementary strengths.

Within the class of pixel-level approaches, unsupervised image translation has been widely explored as a mechanism to bridge visual gaps between domains. Arruda et al. (2022) investigate the use of translation models, including CycleGAN and AST-AdaIN, to adapt synthetic images to the target domain prior to detector training. Their results on the SIM10K $\rightarrow$ Cityscapes benchmark demonstrate that such transformations can improve cross-domain detection performance. However, they also show that improvements in visual appearance do not always translate into optimal task performance, highlighting the limitations of purely image-level adaptation.

In parallel, recent work has extended DA techniques to transformer-based detection architectures. The DATR framework proposed by Chen, Han e Wang (2025) introduces a domain-adaptive detection transformer that combines pseudo-labeling with a mean-teacher strategy and dataset-level adaptation. Additionally, the method incorporates prototypical feature alignment to improve consistency across domains. These results indicate that adaptation strategies can be effectively integrated into modern detection paradigms beyond convolutional networks.

Among the existing approaches, CycleGAN and ConfMix represent two influential and complementary paradigms. CycleGAN addresses domain discrepancies at the pixel level through image-to-image translation, while ConfMix operates at the feature level by leveraging confidence-guided pseudo-labeling and sample mixing. Due to their effectiveness and relevance to Sim2Real object detection, these methods are adopted as the main reference frameworks in this dissertation.

In addition to adaptation strategies applied at the image or feature level, recent work has also explored the role of synthetic data generation as a mean to reduce the domain gap prior to model training. Instead of relying solely on adaptation mechanisms, these approaches aim to construct synthetic datasets that better approximate the characteristics of the target domain.

For example, Marcus et al. (2025) propose a data generation pipeline based on the Car Learning to Act simulator (CARLA) simulator to produce synthetic driving scenes aligned with the KITTI Vision Benchmark Suite (KITTI) dataset. Their approach focuses on controlling scene composition, sensor configuration, and environmental conditions in order to generate data that is structurally and visually closer to real-world driving scenarios. By improving the quality and relevance of synthetic data, this strategy reduces the burden on downstream DA methods.

A closely related line of research focuses on the systematic evaluation of synthetic-to-real transfer through benchmark design. The Synthetic-to-Real benchmark (Syn2Real) benchmark proposed by Peng et al. (2018a) provides a controlled setting to study vi-

sual DA using synthetic Computer-Aided Design (CAD) models and real-world images. By defining standardized tasks and evaluation protocols, this benchmark enables consistent comparison across methods and has contributed to a better understanding of the challenges involved in transferring knowledge from synthetic to real domains.

Together, these works highlight that addressing the Sim2Real gap is not limited to model-level adaptation, but also involves improvements in data generation and evaluation protocols. In this context, synthetic data quality and domain alignment strategies play complementary roles in enabling robust cross-domain generalization.

In addition to approaches that focus on a single level of adaptation, several works have explored hybrid strategies that combine pixel-level and feature-level alignment. These methods aim to jointly reduce visual discrepancies and improve representation consistency across domains.

For instance, Wang et al. (2022) propose the Parallel Teacher framework, which integrates image translation and feature-level adaptation in a unified training scheme. The method leverages CycleGAN-based translation to reduce low-level appearance differences, while a mean-teacher mechanism is used to enforce consistency between teacher and student models during adaptation. Evaluated on the SIM10K $\rightarrow$ Cityscapes benchmark, this approach demonstrates that combining complementary adaptation strategies can lead to improved detection performance compared to using either technique in isolation.

Similarly, SC-UDA by (YU et al., 2022) addresses DA by explicitly modeling both style and content discrepancies between domains. The method combines image translation with pseudo-labeling, aiming to align visual appearance while preserving semantic structure. By distinguishing between style-related variations and content-related information, SC-UDA provides a more structured formulation of the domain gap and shows improved robustness in cross-domain object detection tasks.

These hybrid approaches further reinforce the idea that effective DA often requires the integration of multiple strategies. In particular, they highlight the complementary roles of image translation and feature-level alignment, motivating the investigation of methods that combine both perspectives in a unified framework.

3.2 ConfMix for Feature Level UDA

ConfMix is a recent UDA method for object detection that introduces a novel perspective based on sample mixing guided by detection confidence. Unlike traditional approaches that focus primarily on feature alignment or adversarial learning, ConfMix formulates DA as a data augmentation problem, where synthetic training samples are constructed by combining information from both source and target domains (MATTOLIN et al., 2023).

The method is designed to operate in a standard UDA setting, where a labeled source dataset $\mathcal{D}_S = \{(x_s, y_s)\}$ and an unlabeled target dataset $\mathcal{D}_T = \{x_t\}$ are available. The

central idea is to exploit reliable pseudo-labels generated on the target domain to guide the adaptation process, while preserving the supervision provided by source annotations.

3.2.1 Method Overview

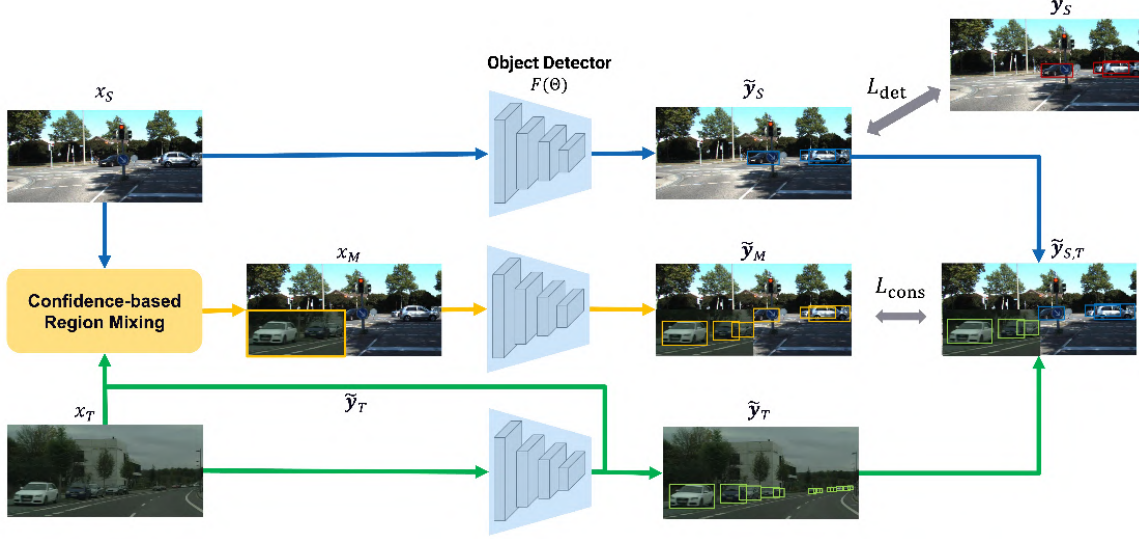


Figure 2 – Overview of the ConfMix framework by Mattolin et al. (2023)

The overall pipeline of ConfMix is illustrated in Figure 2. The method generates a mixed sample x_M by combining a labeled source image x_S with a region from a target image x_T , selected according to pseudo-detection confidence.

Initially, a detector $F(\Theta)$ is trained on the source domain. This model is then used to generate pseudo detections $\tilde{y}_T$ on target images. These predictions are associated with a confidence score that reflects their reliability. The key idea is to identify regions in the target image where the detector is most confident and use these regions to construct mixed samples.

3.2.2 Confidence Estimation

A central component of ConfMix is the definition of a robust confidence metric for pseudo detections. Instead of relying solely on the detector’s classification confidence, the method incorporates uncertainty in bounding box localization.

Each bounding box prediction is modeled as a Gaussian distribution, where both the mean and variance are estimated by the detector. The variance represents the uncertainty associated with the localization. Based on this formulation, a bounding box confidence is computed as:

$$C_{bbx} = 1 - \frac{1}{n} \sum_{i=1}^n \Sigma_i \quad (1)$$

where Σ denotes the predicted variance of the bounding box parameters. This term is combined with the detector confidence C_{det} to produce a final confidence score:

$$C_{comb} = C_{det} \cdot C_{bbx}. \quad (2)$$

This formulation allows the method to prioritize pseudo detections that are both semantically confident and spatially reliable, reducing the impact of noisy predictions.

3.2.3 Confidence-Based Region Mixing

Based on the computed confidence scores, ConfMix introduces a region-level mixing strategy. The target image is divided into multiple regions, and the region with the highest average confidence is selected.

This region is then combined with a source image to generate a new synthetic sample:

$$x_M = M_T \odot x_S + (1 - M_T) \odot x_T, \quad (3)$$

where M_T is a binary mask indicating the selected target region.

This strategy differs from previous approaches such as CutMix by Yun et al. (2019), as it avoids random mixing and instead focuses on semantically meaningful and reliable regions. As a result, the generated samples contain a combination of source annotations and target visual characteristics, facilitating DA.

3.2.4 Progressive Pseudo-Labeling

Another key contribution of ConfMix is the introduction of a progressive pseudo-labeling strategy. The method gradually transitions from a loose confidence criterion to a stricter one during training.

At early stages, the detector uses C_{det} to include a larger number of pseudo detections, allowing the model to explore the target domain. As training progresses, the method shifts towards C_{comb} , which is more restrictive but also more reliable.

This transition is controlled by a weighting factor δ :

$$C = (1 - \delta)C_{det} + \delta C_{comb}, \quad (4)$$

where δ increases over time following a non-linear schedule. This mechanism ensures a smooth transition from exploration to refinement, improving both stability and performance during training.

3.2.5 Training Objective

The training of ConfMix relies on two complementary loss terms: a supervised detection loss and a consistency loss. The supervised term, denoted by L_{det} , is computed on labeled source data using the standard object detection objective. The consistency term, denoted by L_{cons} , is computed on the mixed samples and encourages the detector predictions to agree with pseudo-labels derived from the model’s own predictions on the source and target components of the mixed image.

The total training objective is defined as

$$L_{total} = L_{det} + \gamma L_{cons}, \quad (5)$$

where γ is a dynamic weight that depends on the reliability of the pseudo detections. In this way, the contribution of L_{cons} increases when the pseudo-labels are considered more reliable and decreases when they are more uncertain. This adaptive weighting helps reduce the influence of noisy pseudo-labels and contributes to more stable training.

In practice, L_{det} corresponds to the standard detection loss of the underlying detector, whereas L_{cons} applies the same detection objective to the mixed samples using pseudo-labels derived from the selected target region. The detailed implementation of these terms follows the original formulation of ConfMix (MATTOLIN et al., 2023).

3.2.6 Results Reported by the Authors

ConfMix is evaluated on standard DA benchmarks, including SIM10K $\rightarrow$ Cityscapes and KITTI $\rightarrow$ Cityscapes. The method achieves state-of-the-art performance, outperforming previous approaches by significant margins.

In particular, ConfMix improves detection performance by +1.7% mean Average Precision (mAP) on SIM10K $\rightarrow$ Cityscapes and +3.7% mAP on KITTI $\rightarrow$ Cityscapes compared to prior methods (MATTOLIN et al., 2023).

Additionally, results on the Cityscapes $\rightarrow$ FoggyCityscapes benchmark show that the method approaches the performance of a fully supervised model, with only a small performance gap (MATTOLIN et al., 2023).

Ablation studies further demonstrate that:

- ❑ Confidence-based region selection outperforms random mixing strategies such as CutMix;
- ❑ Progressive pseudo-labeling improves both precision and recall;
- ❑ Dynamic weighting of the consistency loss leads to more stable training.

3.2.7 Discussion

The results indicate that ConfMix provides an effective alternative to traditional feature-alignment methods. By formulating DA as a confidence-guided data mixing problem, the method leverages both source supervision and reliable target information in a unified framework.

One of the main advantages of ConfMix is that it does not require architectural modifications or additional adversarial components, making it compatible with different object detection models. Furthermore, the method is particularly well-suited for one-stage detectors, which are commonly used in real-time applications.

However, the performance of ConfMix still depends on the quality of pseudo detections, especially in early training stages. Although the progressive labeling strategy mitigates this issue, the method may still be sensitive to severe domain shifts.

Overall, ConfMix demonstrates that combining pixel-level information (through mixing) and feature-level learning (through consistency regularization) can lead to robust DA for object detection.

3.3 CycleGAN for Unpaired Image-to-Image Translation

Traditional image-to-image translation methods rely on paired datasets, where corresponding samples across domains are available. However, in many practical scenarios, such paired data is either expensive or infeasible to obtain. This limitation is especially evident in DA tasks such as transferring knowledge from synthetic to real-world environments.

To address this issue, Zhu et al. (2017) proposed CycleGAN, a framework capable of learning mappings between two visual domains without requiring paired examples. The method enables unpaired image-to-image translation by enforcing structural consistency between the input and reconstructed images.

$$G : X \rightarrow Y, \quad F : Y \rightarrow X, \quad (6)$$

$$D_Y : \{y \sim Y\} \cup \{G(x)\}, \quad D_X : \{x \sim X\} \cup \{F(y)\} \quad (7)$$

3.3.1 Model Architecture

CycleGAN consists of two generator networks (G and F) and two discriminator networks (D_Y and D_X). Each generator learns a mapping between domains, while each

discriminator evaluates the realism of generated samples.

The generators follow an encoder–decoder structure with residual blocks in the bottleneck, allowing the model to preserve high-level content while adapting appearance. The discriminators are based on PatchGAN, which classifies overlapping image patches instead of the entire image, encouraging high-frequency realism (ISOLA et al., 2017). The overall CycleGAN architecture is illustrated in Figure 3.

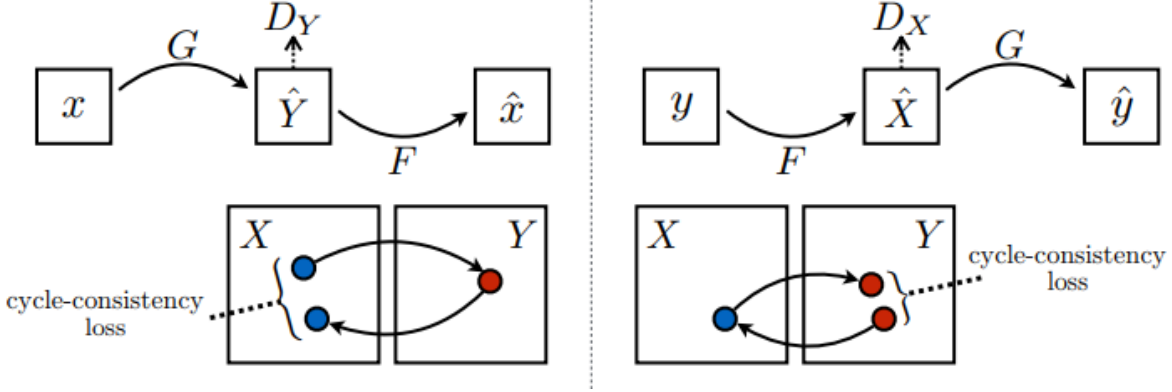


Figure 3 – Overview of the CycleGAN framework by Zhu et al. (2017). Two generators (G and F) learn mappings between domains X and Y , while discriminators (D_X and D_Y) enforce adversarial constraints. Cycle-consistency ensures that translated images can be mapped back to the original domain

3.3.2 Training Objective

The training of CycleGAN relies on three complementary loss terms — an adversarial loss, a cycle consistency loss, and an identity loss — which are combined into a single objective.

The adversarial loss enforces the generated images to match the target domain distribution. For the mapping $G : X \rightarrow Y$, the objective is defined as:

$$\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{\text{data}}(y)} [\log D_Y(y)] + \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log(1 - D_Y(G(x)))]. \quad (8)$$

A similar objective is defined for $F : Y \rightarrow X$.

Since unpaired data does not provide direct supervision, CycleGAN introduces a cycle consistency constraint to regularize the mapping functions. The idea is that translating an image to the target domain and back should recover the original image:

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}}(y)} [\|G(F(y)) - y\|_1]. \quad (9)$$

This loss is critical to prevent arbitrary mappings and ensures structural consistency between domains.

To further preserve color composition and low-level features, an identity mapping loss can be employed:

$$\mathcal{L}_{id}(G, F) = \mathbb{E}_{y \sim p_{data}(y)} [\|G(y) - y\|_1] + \mathbb{E}_{x \sim p_{data}(x)} [\|F(x) - x\|_1]. \quad (10)$$

This term encourages the generators to behave as identity functions when real samples from the target domain are provided as input.

Finally, the complete objective function combines the adversarial, cycle consistency, and identity losses:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G, F) + \gamma_{id} \mathcal{L}_{id}(G, F), \quad (11)$$

where λ and γ_{id} are hyperparameters controlling the relative importance of the cycle consistency and identity loss terms, respectively.

3.3.3 Applications and Limitations in Domain Adaptation

CycleGAN has been widely used in DA scenarios, particularly for bridging the gap between synthetic and real datasets. In this work, it is employed to translate synthetic images into a more realistic domain, aiming to improve object detection performance when training models such as YOLOv5 on translated data.

By reducing the domain discrepancy, the model benefits from both the scalability of synthetic data and the realism of target-domain distributions.

Despite its effectiveness, CycleGAN presents several limitations. The model does not explicitly enforce semantic consistency, which may lead to structural distortions in generated images. For instance, in traffic scenes, the translation may deform vehicle contours, merge nearby cars, or even remove small or distant objects entirely, so that a bounding box inherited from the original synthetic image no longer matches any real object in the translated one. Additionally, training can be unstable due to adversarial optimization, and the method may introduce visual artifacts, such as inconsistent textures on the road surface or hallucinated patterns in the sky and background regions, which do not correspond to any structure present in the original scene. These factors can negatively impact downstream tasks such as object detection if not properly controlled. This limitation is not merely hypothetical: as shown later in the experiments (Section 6.3), in the DOLPHINS→nuScenes scenario the use of translated images degraded detection performance below the source-only baseline, indicating that translation artifacts can actively harm adaptation when structural fidelity is not preserved.

Chapter 4

TransConfMix and ConfMix-Distill

This chapter presents the methods investigated in this dissertation to study the interaction between pixel-level image translation and confidence-guided domain adaptation for object detection in Sim2Real settings. It describes TransConfMix and TransConfMix-2Stage, which incorporate CycleGAN-translated source images into the ConfMix adaptation pipeline, and then introduces ConfMix-Distill, a related variant that augments the adaptation stage with a KL-based consistency regularization term.

Rather than treating translated source images as a beneficial addition to adaptation, the methods introduced in this chapter are designed to examine under which conditions image translation may help or hinder pseudo-labeling-based UDA in one-stage detectors.

The chapter is organized as follows. First, the TransConfMix method is presented in detail, followed by its extension TransConfMix-2Stage. Next, the ConfMix-Distill variant is introduced. Finally, the chapter concludes with brief final remarks.

4.1 TransConfMix

TransConfMix extends the original ConfMix framework by modifying the source data used during the second stage of training. More specifically, instead of using the original labeled synthetic images during adaptation, the method replaces them with their translated counterparts generated by CycleGAN. The confidence-guided adaptation mechanism of ConfMix is otherwise preserved.

The motivation behind this modification stems from a mismatch internal to the ConfMix adaptation stage. During adaptation, the supervised source branch still operates on purely synthetic images, while the pseudo-labeled branch already draws on target-domain data. As a result, the detector is simultaneously exposed to two visually distinct sour-

ces of supervision: synthetic source images and real target regions. By replacing the original synthetic images with their CycleGAN-translated counterparts, TransConfMix reduces this internal appearance gap, so that the supervised source branch becomes visually closer to the target domain. The intuition is that aligning the appearance of both branches may ease the adaptation process, allowing the confidence-guided mixing mechanism to operate on samples that are more visually homogeneous and closer to the target distribution.

ConfMix is adopted in this dissertation as the reference UDA framework for object detection. It follows a two-stage training protocol. In the first stage, a detector is trained using only labeled source images. In the second stage, domain adaptation is performed by combining labeled source samples with pseudo-labeled target samples through a confidence-guided mixing strategy. TransConfMix preserves this general organization and modifies only the source samples used during the adaptation stage.

The original ConfMix pipeline is illustrated in Figure 2. In TransConfMix, the first stage remains unchanged, so the initial detector is still trained on the original synthetic source domain. The modification is introduced only in the second stage.

Let $G_{S \rightarrow T}$ denote the translation function learned by CycleGAN. For each source image x_s , a translated version $x'_s = G_{S \rightarrow T}(x_s)$ is obtained. The original source annotations are then reused, giving rise to a translated source set

$$\mathcal{D}'_S = \{(x'_s, y_s)\}.$$

In the original ConfMix framework, the adaptation stage uses labeled samples from $\mathcal{D}_S$ together with pseudo-labeled target samples. In TransConfMix, this stage is modified by replacing $\mathcal{D}_S$ with $\mathcal{D}'_S$, while keeping the original confidence-guided mixing strategy unchanged. As a result, the mixed samples are constructed from translated source images rather than from the original synthetic ones.

More specifically, the source branch involved in the confidence-guided mixing process is no longer based on the original image x_s , but on its translated counterpart x'_s . Therefore, the method preserves the supervision structure of ConfMix while changing the visual appearance of the source samples used during adaptation.

To make this modification explicit, let $\mathcal{D}_S$ denote the original labeled source dataset, $\mathcal{D}'_S$ the translated source dataset, and $\mathcal{D}_T$ the unlabeled target dataset. The training protocol of ConfMix and TransConfMix can then be summarized as follows:

$$\text{ConfMix :} \quad \text{Stage 1 : } \mathcal{D}_S, \quad \text{Stage 2 : } \mathcal{D}_S + \mathcal{D}_T$$

$$\text{TransConfMix :} \quad \text{Stage 1 : } \mathcal{D}_S, \quad \text{Stage 2 : } \mathcal{D}'_S + \mathcal{D}_T$$

These formulations show that TransConfMix preserves the original adaptation mechanism of ConfMix while modifying the source data used during the second stage.

The same design can be extended in two directions. One possibility is to introduce translated source images already in the first training stage. Another is to preserve the original two-stage organization while adding a consistency regularization term between predictions obtained from original and translated versions of the same source image. These two extensions are described in the following sections.

4.1.1 TransConfMix-2Stage

TransConfMix-2Stage is the first variant considered in this dissertation. It extends the main TransConfMix idea by introducing translated source images not only during adaptation, but also in the initial supervised stage. In this sense, while TransConfMix modifies only the second stage of the original ConfMix pipeline, TransConfMix-2Stage modifies both stages of training.

Let $G_{S \rightarrow T}$ denote the translation function learned by CycleGAN, and let

$$\mathcal{D}'_S = \{(x'_s, y_s)\}$$

be the translated source dataset, where $x'_s = G_{S \rightarrow T}(x_s)$ and the original source annotations y_s are preserved. In TransConfMix-2Stage, the first stage is performed on $\mathcal{D}'_S$ instead of on the original source dataset $\mathcal{D}_S$. The second stage then follows the same strategy adopted in TransConfMix, also using $\mathcal{D}'_S$ during adaptation.

This variant investigates whether the same translated-source strategy becomes more effective when introduced from the beginning of training rather than only during adaptation. By exposing the detector to translated source samples in both stages, TransConfMix-2Stage shifts the supervised branch of the pipeline toward a representation that is visually closer to the target domain throughout the optimization process.

To make the difference between the training strategies explicit, let $\mathcal{D}_S$ denote the original labeled source dataset, $\mathcal{D}'_S$ the translated source dataset, and $\mathcal{D}_T$ the unlabeled target dataset. The corresponding protocols can be summarized as follows:

$$\text{ConfMix :} \quad \text{Stage 1 : } \mathcal{D}_S, \quad \text{Stage 2 : } \mathcal{D}_S + \mathcal{D}_T$$

$$\text{TransConfMix :} \quad \text{Stage 1 : } \mathcal{D}_S, \quad \text{Stage 2 : } \mathcal{D}'_S + \mathcal{D}_T$$

$$\text{TransConfMix-2Stage :} \quad \text{Stage 1 : } \mathcal{D}'_S, \quad \text{Stage 2 : } \mathcal{D}'_S + \mathcal{D}_T$$

These formulations show that both TransConfMix variants preserve the original ConfMix adaptation mechanism while differing in the stage at which translated source samples are introduced. In particular, TransConfMix-2Stage changes not only the adaptation stage, but also the domain in which the detector first learns its supervised representation.

4.2 ConfMix-Distill

ConfMix-Distill is another variant investigated in this dissertation. Unlike TransConfMix-2Stage, which introduces translated source images in both training stages, ConfMix-Distill preserves the original two-stage organization of ConfMix and modifies only the adaptation stage. More specifically, the method augments the original objective with a Kullback–Leibler (KL) divergence term computed between predictions obtained from the original and translated versions of the same source sample.

This consistency constraint is applied only during the second stage of the two-stage training protocol. The first stage remains unchanged and is used to obtain the initial detector that later defines the teacher model. During adaptation, the method follows a teacher–student formulation in which two instances of the same YOLOv5 detector are used: a teacher model and a student model. The teacher operates in evaluation mode and remains fixed during training, while the student, initialized from this same first-stage model, is optimized through backpropagation. Keeping the teacher fixed ensures that the consistency target remains stable throughout adaptation, anchoring the regularization to the source-domain knowledge acquired during the first stage; if both predictions were produced by the same continuously updated model, the reference itself would drift during training. Therefore, the proposed method does not modify the detector architecture itself, but augments the training objective used during adaptation.

This variant is motivated by the premise that while image translation reduces holistic domain discrepancy (e.g., global illumination and texture), it may concurrently introduce localized semantic distortions. Enforcing consistency aims to regularize the detector against such translation-induced artifacts. Even when the semantic content of the scene is preserved, the model may respond differently to the original source image and to its translated counterpart. This sensitivity is especially relevant during adaptation, where training already depends on pseudo-label reliability and confidence-based learning.

Based on this observation, the proposed regularization assumes that the detector should produce consistent predictions for different representations of the same source sample. If an original synthetic image x_s and its translated version x'_s depict the same scene, their objectness predictions should remain close despite differences in color, texture, and other low-level cues. Enforcing this consistency may reduce sensitivity to translation-induced variations and lead to a more stable adaptation process.

The proposed regularization is applied only during the second stage of the two-stage training protocol. The first stage remains unchanged and is used to obtain the initial detector that later defines the teacher model. During adaptation, the method follows a teacher–student formulation in which two instances of the same YOLOv5 detector are used: a teacher model and a student model. The teacher operates in evaluation mode and remains fixed during training, while the student is optimized through backpropagation. Keeping the teacher fixed ensures that the consistency target remains stable throughout

adaptation, anchoring the regularization to the source-domain knowledge acquired during the first stage; if both predictions were produced by the same continuously updated model, the reference itself would drift during training. Therefore, the proposed method does not modify the detector architecture itself, but augments the training objective used during adaptation.

The teacher is defined by the detector obtained after the first stage of ConfMix, while the student is initialized from this same model and then optimized during adaptation. For each source sample, the teacher receives the original synthetic image x_s , whereas the student receives its translated counterpart x'_s , generated by the CycleGAN model selected for the corresponding domain pair. Since x_s and x'_s correspond to the same scene, they are treated as spatially aligned counterparts that differ primarily in visual appearance.

During the adaptation stage, the translated source image x'_s is used not only for the KL-based consistency constraint, but also in the supervised source branch of training. In practice, this means that the student is trained on translated source samples for the detection loss and that these translated samples also replace the original source images in the confidence-guided mixing process used to compute the consistency loss. In contrast, the original source image x_s is used exclusively as input to the teacher branch for the distillation term.

The overall structure of the proposed regularized adaptation pipeline is illustrated in Figure 4.

In addition to the original supervised and consistency-based losses of ConfMix, ConfMix-Distill introduces a consistency constraint between predictions obtained from the original and translated versions of the same source image. In this way, the method investigates whether translation-aware prediction consistency can improve robustness during domain adaptation.

Let $p_T^{b,i,j,a,s}$ and $p_S^{b,i,j,a,s}$ denote the sigmoid-activated objectness probabilities predicted by the teacher and student models, respectively, for batch element b , grid cell (i, j) , anchor a , and detection scale s . These probabilities can be interpreted as Bernoulli distributions over object presence. The KL divergence is computed element-wise over all objectness predictions in order to encourage the student outputs to match the teacher predictions.

Formally, the regularization term is defined as

$$L_{KL} = \frac{1}{S} \sum_{s=1}^S \left[\frac{1}{N_s} \sum_{b,a,i,j} D_{\text{KL}}(p_T^{b,i,j,a,s} \parallel p_S^{b,i,j,a,s}) \right], \quad (12)$$

where S denotes the number of detection scales and N_s is the total number of objectness predictions at scale s . This formulation gives equal weight to each detection scale by first averaging the divergence over all predictions within a scale and then averaging the resulting values across scales.

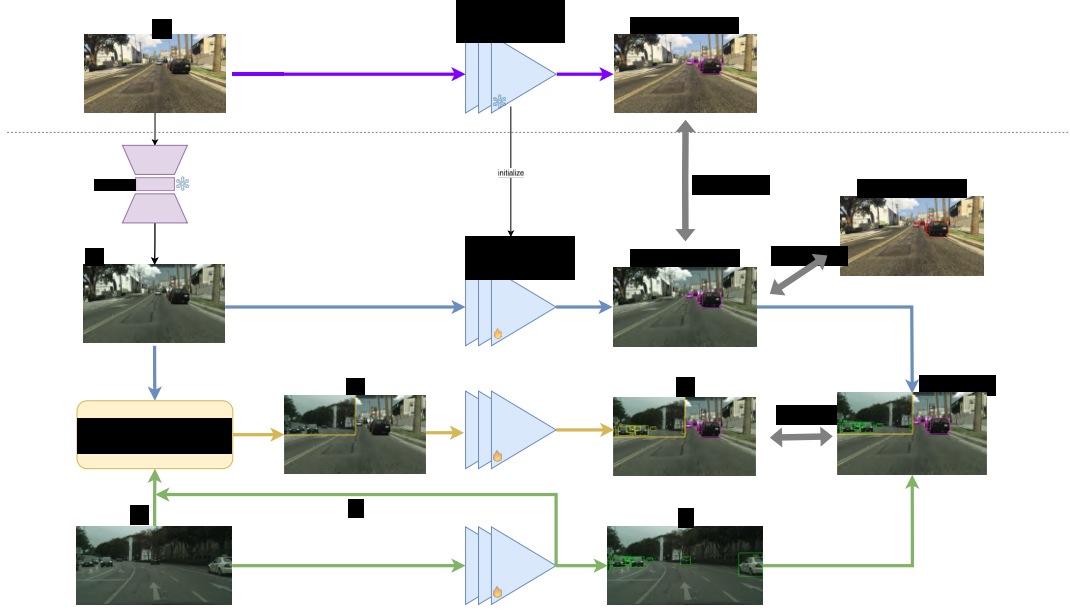


Figure 4 – Overview of the proposed KL-regularized adaptation pipeline (ConfMix-Distill), highlighting the modifications introduced to the original ConfMix framework. The pipeline incorporates translated source images and an additional consistency constraint between predictions. The original synthetic image x_s and its translated counterpart x'_s (generated by CycleGAN) are processed by two detectors: a teacher model $F^T(\theta)$, obtained after the first stage of ConfMix, and a student model $F(\theta)$ initialized from the teacher. The teacher produces predictions $\tilde{y}_s$ from x_s , while the student produces $\tilde{y}'_s$ from x'_s . A KL divergence term L_{KL} is computed between these predictions to enforce consistency across visual domains. In parallel, the student is trained using the ground-truth annotations y_s through the detection loss L_{det} applied to $\tilde{y}'_s$. The adaptation process also includes the original ConfMix consistency term L_{cons} , computed using mixed samples. The final objective combines all loss components

Under the Bernoulli interpretation of the objectness predictions, the divergence at each prediction location is given by

$$D_{\text{KL}}(p_T^{b,i,j,a,s} \parallel p_S^{b,i,j,a,s}) = p_T^{b,i,j,a,s} \log \left(\frac{p_T^{b,i,j,a,s}}{p_S^{b,i,j,a,s}} \right) + (1 - p_T^{b,i,j,a,s}) \log \left(\frac{1 - p_T^{b,i,j,a,s}}{1 - p_S^{b,i,j,a,s}} \right). \quad (13)$$

Importantly, this regularization is applied only to the objectness predictions, without directly constraining the class probabilities or the bounding-box regression outputs. In the implementation, the probabilities are also clamped to a small interval around (0, 1) for numerical stability before computing the divergence.

The final training objective is defined as:

$$L_{total} = L_{det} + \gamma L_{cons} + \lambda_{KL} L_{KL} \quad (14)$$

where L_{det} denotes the supervised detection loss, L_{cons} is the consistency-based adaptation loss inherited from ConfMix, and L_{KL} is the KL divergence regularization term. The factor γ controls the contribution of the consistency loss and is computed dynamically during training from the proportion of pseudo-labels above the confidence threshold, while λ_{KL} weights the influence of the proposed regularization.

By enforcing consistency between predictions obtained from original and translated versions of the same source image, this formulation seeks to reduce sensitivity to appearance changes introduced by image translation. ConfMix-Distill therefore extends ConfMix with a translation-aware regularization term while preserving its original two-stage adaptation structure.

4.3 Final Remarks

This chapter presented TransConfMix together with two related extensions built on top of the original ConfMix framework. All three methods preserve the confidence-guided adaptation mechanism of ConfMix as the basis for simulation-to-real transfer, while modifying specific components of the training process. Table 1 summarizes the four methods side by side, making explicit that they differ in only two aspects: which source images — original ($\mathcal{D}_S$) or translated ($\mathcal{D}'_S$) — are used in each training stage, and whether the additional KL-based consistency term is applied.

Table 1 – Comparison of the training protocols of ConfMix and the proposed variants. In all cases the second stage also uses the unlabeled target data $\mathcal{D}_T$. The last column indicates whether the KL-based consistency term is added.

Method	Stage 1	Stage 2	KL term
ConfMix	$\mathcal{D}_S$	$\mathcal{D}_S + \mathcal{D}_T$	—
TransConfMix	$\mathcal{D}_S$	$\mathcal{D}'_S + \mathcal{D}_T$	—
TransConfMix-2Stage	$\mathcal{D}'_S$	$\mathcal{D}'_S + \mathcal{D}_T$	—
ConfMix-Distill	$\mathcal{D}_S$	$\mathcal{D}'_S + \mathcal{D}_T$	✓

TransConfMix modifies the second stage of ConfMix by replacing the original synthetic source images with their translated counterparts during adaptation. This design investigates whether bringing the supervised source branch closer to the target visual domain at the image level can improve the effectiveness of confidence-guided adaptation. TransConfMix-2Stage extends this idea by introducing translated source images in both stages of training, thereby shifting the supervised branch toward the target style from the beginning of optimization. ConfMix-Distill, in turn, augments the adaptation stage with a KL-based consistency term between predictions obtained from original and translated versions of the same source image, while also using translated source samples in the student branch during adaptation. In particular, ConfMix-Distill can be seen as TransConfMix augmented with the KL consistency term: the student branch follows the same

translated-source configuration ($\mathcal{D}'_S$) used in TransConfMix, while the original source image x_s is additionally routed to a frozen teacher to provide the regularization target.

Taken together, these methods explore two complementary directions for extending ConfMix. The first consists of modifying the source data used during training through image translation, while the second introduces an additional regularization mechanism during adaptation. In all cases, however, the detector architecture remains unchanged, and the original ConfMix formulation serves as the methodological foundation of the proposed approach.

The next chapter describes the experimental methodology adopted to evaluate these methods under controlled simulation-to-real adaptation scenarios.

Chapter 5

Methodology

This chapter describes the experimental methodology adopted to evaluate DA strategies for Sim2Real object detection. The study focuses on a setting in which object detectors are trained using labeled synthetic data and evaluated on real-world images, without access to target-domain annotations during adaptation.

The evaluation protocol is designed to analyze how different adaptation strategies affect cross-domain generalization. In particular, the methodology considers pixel-level and feature-level approaches, as well as their combination, under a controlled setup.

All methods are evaluated under consistent training conditions. Multiple source-target combinations are considered to assess robustness across different types of domain shift.

This chapter is organized as follows. First, the datasets and DA scenarios are presented. Next, the training configuration and reference models are described. The chapter then details the CycleGAN model selection procedure, the compared methods, and the evaluation metrics used to assess detection performance.

5.1 Datasets

The experiments in this dissertation use both synthetic and real-world datasets commonly adopted in object detection and autonomous driving research. The synthetic datasets serve as labeled source domains, while the real-world datasets are used as target domains for evaluation. This setting reflects a typical Sim2Real adaptation scenario, in which labeled simulated data is available whereas target-domain annotations are not used during adaptation, except in the oracle setting.

In order to ensure consistency across datasets, only the vehicle class is considered in all experiments. This choice allows label compatibility between synthetic and real-world

domains and avoids inconsistencies arising from differences in annotation taxonomies. In addition, only front-facing camera images are used for every dataset, all captured from the point of view of a vehicle. This ensures that the source and target domains share a comparable acquisition geometry, isolating the domain shift to differences in visual appearance and scene content rather than to camera viewpoint.

5.1.1 Synthetic Datasets

Two synthetic datasets are used as source domains: DOLPHINS and SIM10K.

DOLPHINS: DOLPHINS is a synthetic autonomous driving dataset generated in the CARLA simulator and originally proposed for collaborative perception research (MAO et al., 2022). In this work, camera images from DOLPHINS are used as a synthetic source domain for vehicle detection. The dataset contains 7,538 images with a resolution of 1920×1080 pixels. Because the data is generated in a simulated environment, bounding box annotations are automatically available. Only front-view camera images, captured from the ego-vehicle perspective, are considered. An example image is shown in Fig. 5.



Figure 5 – Example image from the DOLPHINS dataset used as a synthetic source domain

SIM10K: SIM10K is a synthetic dataset derived from images rendered in the Grand Theft Auto V game engine, containing 10,000 images with bounding box annotations for car detection tasks (JOHNSON-ROBERSON et al., 2016). The images have a resolution of approximately 1912×1052 pixels. In this work, SIM10K is used entirely as a labeled source domain for training, since evaluation is performed on the target domains. No explicit validation split is required for this dataset in the adopted experimental protocol. An example image is shown in Fig. 6.



Figure 6 – Example image from the SIM10K dataset used as a synthetic source domain

5.1.2 Real-World Datasets

Two real-world datasets are used as target domains for evaluation: Cityscapes and nuScenes.

Cityscapes: Cityscapes is a large-scale dataset designed for urban scene understanding in autonomous driving applications (CORDTS et al., 2016). It contains 3,475 high-resolution images (2048×1024 pixels) collected in several European cities. In this work, 2,975 images are used for training, while the remaining images are used for evaluation. Only the vehicle class is considered. An example image is shown in Fig. 7.



Figure 7 – Example image from the Cityscapes dataset used as a real-world target domain

nuScenes: nuScenes is a large-scale autonomous driving dataset that includes multimodal sensor data such as cameras, Light Detection and Ranging (LiDAR), and radar (CAESAR et al., 2020). In this work, only camera images are used, specifically the front-facing view (CAM_FRONT). The dataset contains 33,299 images with a resolution of 1600×900 pixels. From these, 27,430 images are used for training, while the remaining images are used for evaluation. As in the other datasets, only the vehicle class is

considered. An example image is shown in Fig. 8.



Figure 8 – Example image from the nuScenes dataset used as a real-world target domain

5.1.3 Summary of Datasets

Table 2 summarizes the number of annotated vehicle instances in each dataset considered in this dissertation. These values provide an additional characterization of the experimental data beyond the number of images and help illustrate differences in dataset scale and object density across domains.

Table 2 – Number of annotated vehicle instances in each dataset

Dataset	Vehicle Instances
DOLPHINS	30,522
SIM10K	57,776
Cityscapes	38,679
nuScenes	184,353

As can be observed, the datasets differ substantially in the total number of annotated vehicle instances. In particular, nuScenes contains a considerably larger number of vehicle annotations than the other datasets, while DOLPHINS provides the smallest total among the four datasets. Such differences are relevant in the context of Sim2Real transfer, as they reflect variations in scene composition, annotation density, and domain complexity.

5.2 Domain Adaptation Scenarios

To evaluate the domain adaptation methods considered in this dissertation, this work defines multiple Sim2Real adaptation scenarios. In all cases, synthetic datasets are used as labeled source domains, while real-world datasets are treated as unlabeled target domains during adaptation and used for evaluation.

The source domains are DOLPHINS and SIM10K, both providing automatically generated bounding box annotations for the vehicle class. The target domains are Cityscapes and nuScenes, which consist of real urban driving scenes with distinct visual characteristics.

By combining the source and target datasets, four DA scenarios are defined:

- DOLPHINS $\rightarrow$ Cityscapes
- DOLPHINS $\rightarrow$ nuScenes
- SIM10K $\rightarrow$ Cityscapes
- SIM10K $\rightarrow$ nuScenes

Each scenario represents a distinct type of domain shift, involving differences in rendering realism, scene composition, and environmental conditions. DOLPHINS corresponds to a simulation-based synthetic environment, whereas SIM10K is derived from a game engine with a distinct visual style. Similarly, Cityscapes and nuScenes are real-world datasets with different geographic locations, sensor characteristics, and scene distributions.

Evaluating adaptation methods across these domain pairs enables a systematic analysis of cross-domain generalization under varying domain gaps.

5.3 Training Configuration

All object detection models in the experiments are based on the YOLOv5 architecture. A consistent training protocol is used across all methods to ensure comparability.

The choice of YOLOv5 was made primarily to preserve reproducibility and to ensure a fair comparison with the original ConfMix framework, which is also built on YOLOv5. Although more recent detector architectures are available, using the same base model as the reference method avoids introducing additional confounding factors related to architectural differences. In this way, the experimental comparison remains centered on the adaptation strategies themselves rather than on gains that could arise from changes in the underlying detector.

The experimental setup is designed to isolate the effect of the DA strategy. Therefore, all models share the same detector architecture, optimizer configuration, and data augmentation procedures, unless a specific method requires otherwise.

Two training paradigms are considered in this work: single-stage supervised training and two-stage domain adaptation. The reference models, namely Baseline and Oracle, follow a standard single-stage supervised training protocol for 100 epochs. In contrast, adaptation-based methods follow a two-stage procedure. The first stage consists of source-domain training for 50 epochs, during which the detector is trained using only labeled

synthetic data. The second stage performs UDA for an additional 50 epochs using unlabeled target-domain images together with source-domain supervision, according to the adaptation strategy of each method.

In all experiments, the detector from the final training epoch is used for evaluation instead of selecting the best validation checkpoint. This decision follows directly from the unsupervised nature of the adaptation setting. Selecting the best checkpoint requires a labeled validation set, but in UDA no annotations are available in the target domain; using such a checkpoint would therefore require access to target labels, violating the core assumption of the scenario and introducing information leakage from the evaluation domain into model selection. Relying on a source-domain validation set would not solve this issue either, since performance on the synthetic source is not a reliable indicator of performance on the real target under domain shift. Using the detector from the final epoch avoids this dependency entirely, at the cost of not necessarily corresponding to the single best-performing checkpoint for every method. This choice avoids bias from target-domain validation data and ensures a consistent evaluation protocol across methods.

Unless otherwise specified, the same training configuration is maintained throughout the experiments, so that performance differences can be attributed to the adaptation strategy rather than to variations in training conditions.

5.4 Reference Models

To establish reference performance levels, two models are considered throughout the experiments: a baseline detector and an oracle detector.

The **baseline** model represents the performance obtained when training is performed exclusively on labeled synthetic data. In this setting, a YOLOv5 detector is trained on the source domain and directly evaluated on the target domain without any DA. This model defines a lower-bound reference for cross-domain generalization under domain shift.

The **oracle** model represents an upper-bound reference, obtained under the assumption that labeled target-domain data is available during training. In this case, the detector is trained on the target domain and evaluated on its corresponding test set, effectively removing the domain shift between training and testing data. It is important to note that this setting does not correspond to the UDA scenario, but is included solely as a performance reference.

Both reference models follow the same detector architecture and training configuration, differing only in the domain used for supervision. Together, they define a performance interval that allows quantifying how close each adaptation method approaches the ideal target-domain performance.

5.5 CycleGAN Hyperparameter Search

CycleGAN-based image translation is used as a component of the evaluated adaptation strategies. Since the quality of the translated images can significantly affect downstream object detection performance, a model selection procedure is conducted to determine, for each DA scenario, the most effective translation configuration.

For each source-target pair defined in Section 5.2, four CycleGAN configurations are trained using different hyperparameter settings. The explored configurations vary along three axes: the image preprocessing, the learning-rate policy, and the use of the identity-loss term. For image preprocessing, the configurations either apply explicit resizing and cropping (`load_size` = 1024, `crop_size` = 512) or use no explicit resizing or cropping. For the learning-rate policy, they alternate between a constant schedule and a plateau-based schedule. For the identity-loss term, they either include it (`lambda_identity` = 1.0) or omit it entirely. All remaining settings are kept fixed across configurations, including 100 epochs of initial training followed by 100 epochs of learning-rate decay, a batch size of 1, a learning rate of 0.0001, and no dropout. The four resulting configurations are described in detail in Appendix A, and the same four are evaluated for every adaptation pair.

After training each CycleGAN model, the entire synthetic source dataset is translated to the visual style of the corresponding target domain. To evaluate the effectiveness of each configuration, a task-driven criterion is adopted. Specifically, for each translated dataset, a YOLOv5 detector is trained for 50 epochs using the original synthetic annotations.

The trained detector is then evaluated on the corresponding target domain, and the resulting detection performance is used as a proxy measure for translation quality. This approach ensures that the selection of the CycleGAN configuration is directly aligned with the final object detection objective, rather than relying on qualitative visual inspection.

The best configuration for each domain pair is selected based on the highest detection performance on the target domain, considering both $\text{mAP}@0.5$ and $\text{mAP}@[0.5:0.95]$. This selection procedure is applied independently for each adaptation scenario, resulting in a specific translation model tailored to each source-target combination.

5.6 Compared Methods

The experiments compare reference models, existing DA strategies, and the variants introduced in Chapter 4 in order to analyze the impact of different approaches to Sim2Real transfer. In particular, the comparison is designed to isolate the contributions of pixel-level adaptation, feature-level adaptation, and their combination.

The evaluated methods are grouped as follows:

□ Reference Models

- **Baseline:** detector trained exclusively on labeled synthetic source data and evaluated on the real target domain without adaptation, providing a lower-bound reference.
- **Oracle:** detector trained on labeled target-domain data and evaluated on the corresponding test set, providing an upper-bound reference.

□ Pixel-Level Adaptation

- **CycleGAN:** detector trained on translated synthetic images using the selected CycleGAN configuration.

□ Feature-Level Adaptation

- **ConfMix:** feature-level DA method based on confidence-guided mixing.

□ Proposed Variants

- **TransConfMix:** proposed variant using translated source images during the adaptation stage.
- **TransConfMix-2Stage:** proposed variant using translated images in both training stages.
- **ConfMix-Distill:** proposed variant with KL-based consistency regularization, evaluated with three values of the regularization weight, $\lambda_{KL} \in \{0.001, 0.1, 0.5\}$. These values were selected to probe the behavior of the regularization across distinct orders of magnitude, ranging from a near-negligible contribution ($\lambda_{KL} = 0.001$) to a moderate one ($\lambda_{KL} = 0.5$), rather than to perform an exhaustive hyperparameter search.

5.7 Evaluation Metrics

The performance of all object detection models is evaluated using mAP metrics based on Intersection over Union (IoU). These metrics assess detection quality by jointly considering localization accuracy and classification correctness.

The evaluation is performed on the target-domain test set for all methods, ensuring a consistent comparison under the UDA setting.

Two evaluation metrics are used in this work:

- **mAP@0.5:** mAP computed at a fixed IoU threshold of 0.5. A predicted bounding box is considered a true positive if its IoU with a ground-truth box exceeds this threshold. This metric is commonly associated with the PASCAL VOC evaluation protocol and provides a less strict assessment of localization quality.

- **mAP@[0.5:0.95]**: mAP averaged over IoU thresholds ranging from 0.5 to 0.95 in increments of 0.05. This COCO-style metric provides a more demanding evaluation by requiring accurate localization across multiple IoU thresholds.

All methods are evaluated under the same metric computation protocol, ensuring that performance differences are attributable to the adaptation strategy rather than to variations in evaluation procedures.

Using both metrics enables a more comprehensive assessment of detection performance. While mAP@0.5 captures general detection capability under a permissive threshold, mAP@[0.5:0.95] emphasizes precise localization and provides a stricter evaluation of model performance.

Chapter 6

Results and Analysis

This chapter presents the experimental results obtained in the Sim2Real DA experiments. The objective of the evaluation is to analyze the impact of different adaptation strategies on object detection performance when training is performed using synthetic data and evaluation is conducted on real-world datasets.

The chapter is organized into four main parts. First, the impact of different CycleGAN configurations is evaluated in order to determine the most effective image translation settings for each domain pair. Next, the detection results obtained with the selected CycleGAN configurations are presented. Then, the performance of DA methods is compared, including ConfMix and the proposed TransConfMix variants. Finally, the effect of the proposed KL-based regularization term is analyzed.

6.1 CycleGAN Hyperparameter Evaluation

As described in Chapter 5, CycleGAN is used as a pixel-level adaptation component to reduce the visual discrepancy between synthetic and real domains. Since the effectiveness of this step depends strongly on the translation configuration, a systematic evaluation of multiple CycleGAN settings was conducted for each source–target pair.

Instead of assessing translation quality solely through visual inspection, a task-driven evaluation strategy was adopted. For each configuration, the synthetic source dataset was translated to the target domain, and a YOLOv5 detector was trained on the translated images using the original annotations. The resulting detection performance on the target dataset was then used as a proxy measure for translation quality. Complete results for all evaluated configurations are provided in Appendix A.

For reference, the four evaluated configurations differ along three axes — image preprocessing, learning-rate policy, and identity-loss term — and can be summarized as follows. Configuration 1 applies explicit resizing and cropping (`load_size` = 1024, `crop_size` = 512) with a constant learning-rate schedule and identity loss enabled (`lambda_identity` = 1.0). Configuration 2 keeps the constant schedule and identity loss but uses no explicit resizing or cropping. Configuration 3 matches the preprocessing of Configuration 1 but replaces the constant schedule with a plateau-based one. Configuration 4 uses no explicit resizing or cropping, a plateau schedule, and no identity-loss term. The complete settings and per-configuration results are provided in Appendix A.

Table 3 summarizes the best-performing configuration selected for each domain pair according to detection performance.

Table 3 – Selected CycleGAN configurations for each domain adaptation scenario

Domain Pair	Configuration	mAP@0.5	mAP@[0.5:0.95]
SIM10K → Cityscapes	Config 3	0.3814	0.2301
DOLPHINS → Cityscapes	Config 1	0.1698	0.0864
SIM10K → nuScenes	Config 4	0.3835	0.1543
DOLPHINS → nuScenes	Config 4	0.2635	0.1388

The results indicate that the optimal CycleGAN configuration is not consistent across domain pairs. Different source-target combinations require distinct translation settings, reflecting variations in appearance gap, scene structure, and dataset characteristics.

In particular, configurations that modify the learning-rate scheduling and identity-loss weighting show different impacts depending on the domain pair. This suggests that the effectiveness of pixel-level adaptation is sensitive not only to the translation model itself but also to the specific properties of the domains involved. Concretely, three of the four domain pairs selected configurations using the plateau schedule (Configurations 3 and 4), and both nuScenes targets selected Configuration 4, in which the identity-loss term is disabled — suggesting that, for these pairs, relaxing the identity constraint allowed the translation to move further toward the target appearance.

The selected configurations are used in all subsequent experiments, ensuring that the evaluation of DA methods is based on the most effective translation setup for each case.

Figure 9 provides qualitative examples of the translated images obtained with the selected configurations. In general, the translations modify color, contrast, and texture statistics in the direction of the target domain, but the visual fidelity is not uniform across source-target pairs. In the DOLPHINS → nuScenes case, the translated image exhibits more pronounced artifacts, particularly in background regions such as the sky, buildings, and other large scene elements. These degradations suggest that the translation process does not affect only low-level appearance, but may also alter contextual structure and introduce unrealistic patterns around the objects of interest. Although the vehicles

remain visible, the surrounding scene becomes visually less coherent, which may weaken the detector training signal.

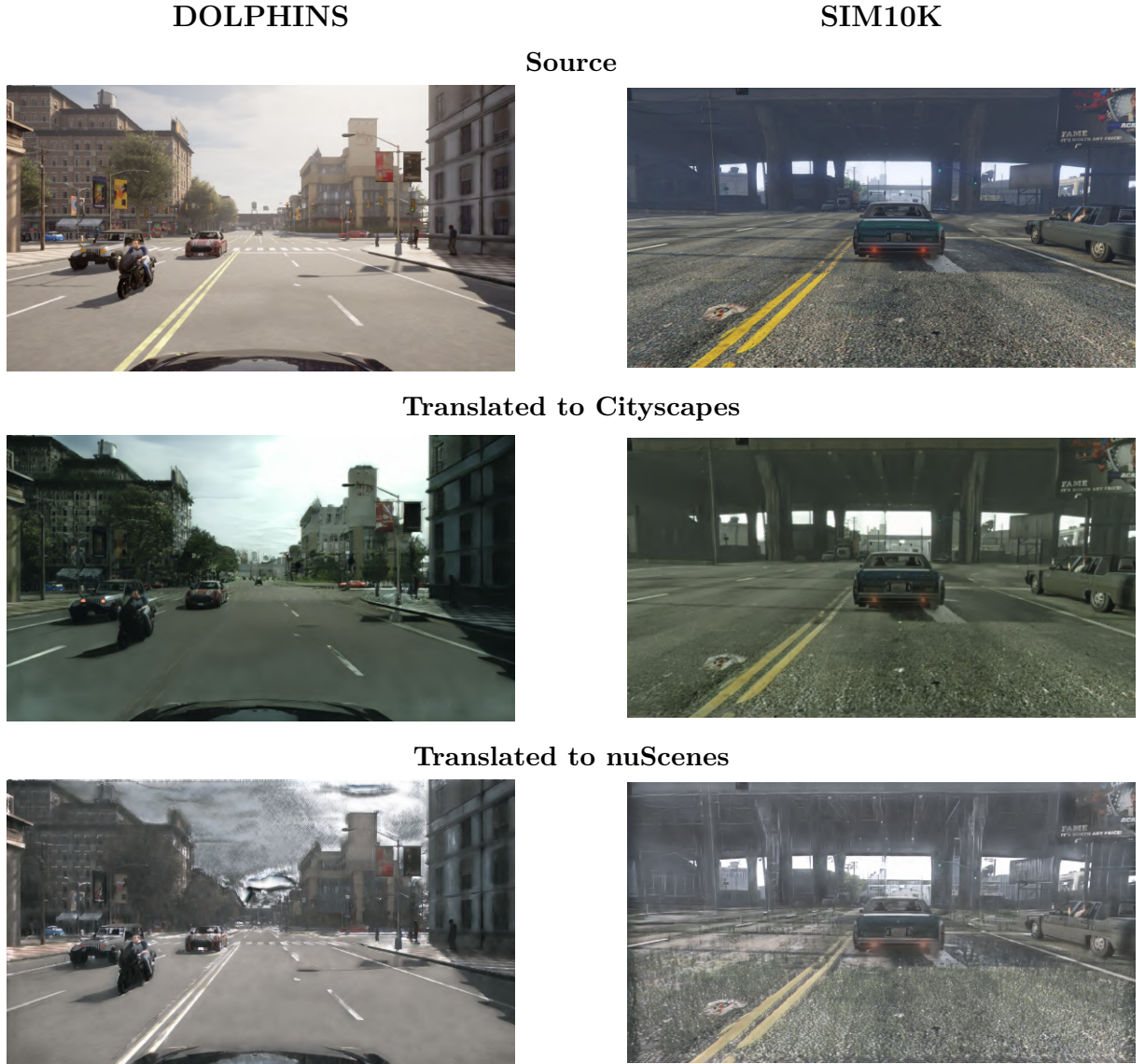


Figure 9 – Qualitative comparison of CycleGAN-based image translation. The two columns correspond to the DOLPHINS and SIM10K source datasets, while each row shows the original image and its translated versions in the Cityscapes and nuScenes domains

6.2 Comparison of Domain Adaptation Methods

This section compares the object detection results obtained with the evaluated DA methods across the four Sim2Real scenarios considered in this dissertation. The comparison includes the reference models, the original ConfMix method, the proposed Trans-ConfMix variants, and the KL-regularized ConfMix.

For each domain pair, the results will be reported using $\text{mAP}@0.5$ and $\text{mAP}@[0.5:0.95]$, allowing a direct comparison between source-only training, translated-image training, feature-level adaptation, and the proposed extensions.

Before presenting the per-scenario results, it is worth examining the specific mechanisms through which translation defects can propagate into the detector and contaminate its training signal. Because TransConfMix and ConfMix-Distill reuse the original bounding-box annotations on the translated images, any geometric change introduced by CycleGAN breaks the alignment between a label and the object it is supposed to enclose. Two failure modes are particularly relevant. First, spatial displacement: when the translation slightly shifts, rescales, or deforms a vehicle, its real position no longer coincides with the inherited bounding box, so the detector is trained to associate the object with a misplaced target, degrading localization accuracy — an effect consistent with the lower $\text{mAP}@[0.5:0.95]$ observed for translation-based methods in several scenarios. Second, object-boundary artifacts: CycleGAN may blur, duplicate, or hallucinate edges around vehicles, or blend them into the background, which corrupts precisely the contours the detector relies on to regress box coordinates and to estimate objectness. When such artifacts are severe, the supervised loss computed on translated samples reinforces incorrect appearance-label associations rather than correcting them. These two mechanisms help explain why methods that depend on translated images do not consistently improve over ConfMix, and why their advantage is greatest in scenarios where the translation preserves structural fidelity and smallest — or even negative — where it does not.

6.2.1 SIM10K to Cityscapes

Table 4 – Detection performance for the SIM10K $\rightarrow$ Cityscapes adaptation scenario

Method	$\text{mAP}@0.5$	$\text{mAP}@[0.5:0.95]$
Baseline	0.3172	0.1998
CycleGAN	0.3630	0.2196
ConfMix	0.4082	0.2396
TransConfMix	0.3258	0.1924
TransConfMix-2Stage	0.3904	0.2286
ConfMix-Distill ($\lambda_{KL} = 0.5$)	0.3955	0.2181
ConfMix-Distill ($\lambda_{KL} = 0.1$)	0.3698	0.2038
ConfMix-Distill ($\lambda_{KL} = 0.001$)	0.3704	0.2124
Oracle	0.5824	0.3817

Table 4 presents the detection results for the SIM10K $\rightarrow$ Cityscapes adaptation scenario. All evaluated DA methods improved performance over the source-only baseline, confirming the relevance of adaptation strategies in reducing the Sim2Real gap.

Among the evaluated methods, ConfMix achieved the best performance, indicating that feature-level adaptation guided by confidence-based pseudo-labeling is particularly

effective in this scenario. This suggests that reliable target-domain predictions provide a stronger adaptation signal than pixel-level transformations alone.

CycleGAN-based training also led to performance gains compared to the baseline, suggesting that the translation step reduces part of the low-level visual discrepancy between the synthetic and real domains. In particular, changes in texture statistics, color distribution, and overall image appearance may make the translated source images more similar to the target data, which improves the usefulness of source supervision. However, its performance remains below that of ConfMix, indicating that appearance alignment alone is not sufficient to fully address the domain shift. A likely reason is that pixel-level translation does not explicitly exploit target-domain predictions during training and may still preserve residual structural mismatches that are better handled by confidence-guided feature-level adaptation.

The TransConfMix variants produced mixed results. While the two-stage version achieved competitive performance, the single-stage TransConfMix underperformed compared to both ConfMix and CycleGAN. One possible interpretation is that replacing the original source images with translated ones during adaptation may alter the supervision conditions of the source branch in a way that is not always beneficial. In particular, when the translated images contain artifacts or imperfect structural preservation, the resulting source supervision may become less reliable, which can negatively affect the adaptation process.

The KL-regularized variants consistently improved over the baseline but did not surpass the original ConfMix model. Among the tested configurations, $\lambda_{KL} = 0.5$ achieved the best performance within this group, indicating that a moderate regularization strength is beneficial. However, stronger reliance on the distillation signal may constrain the learning process, limiting the model’s ability to adapt to target-domain characteristics.

Figure 10 presents a qualitative comparison of detection results for the TransConfMix variants and the KL-regularized models. The examples highlight differences in localization quality and confidence calibration across methods, particularly in regions containing multiple vehicles and partial occlusions.

6.2.2 DOLPHINS to Cityscapes

Table 5 presents the results for the DOLPHINS $\rightarrow$ Cityscapes adaptation scenario. Compared to the SIM10K-based experiments, this setting is more challenging, as reflected by the significantly lower baseline performance.

All evaluated adaptation strategies improved over the source-only baseline, confirming the effectiveness of DA in this scenario. In particular, CycleGAN-based training led to a substantial relative gain, indicating that reducing low-level visual discrepancies is especially beneficial when the initial domain gap is large.

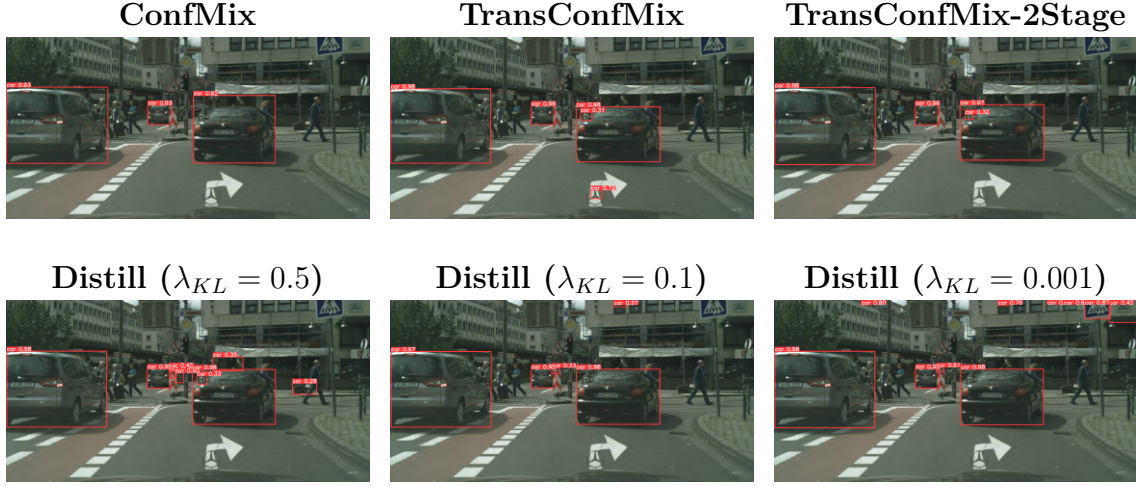


Figure 10 – Qualitative comparison of detection results for the SIM10K → Cityscapes scenario. Each column corresponds to a different adaptation method, showing representative predictions on the same target-domain image

Table 5 – Detection performance for the DOLPHINS → Cityscapes adaptation scenario

Method	mAP@0.5	mAP@[0.5:0.95]
Baseline	0.1053	0.0481
CycleGAN	0.1642	0.0927
ConfMix	0.2883	0.1413
TransConfMix	0.2591	0.1329
TransConfMix-2Stage	0.2494	0.1185
ConfMix-Distill ($\lambda_{KL} = 0.5$)	0.2428	0.1181
ConfMix-Distill ($\lambda_{KL} = 0.1$)	0.2717	0.1355
ConfMix-Distill ($\lambda_{KL} = 0.001$)	0.2693	0.1243
Oracle	0.5824	0.3817

Among the evaluated methods, ConfMix achieved the best overall performance, demonstrating that feature-level adaptation remains effective even under more challenging domain shifts. The large improvement over the baseline suggests that confidence-guided pseudo-labeling provides a strong supervisory signal when transferring from highly synthetic data.

The TransConfMix variants also achieved competitive results, with the single-stage version outperforming its two-stage counterpart in this scenario. This behavior differs from the SIM10K → Cityscapes case and suggests that the effectiveness of incorporating translated images depends on the characteristics of the source domain.

The KL-regularized variants consistently improved over the baseline but did not surpass ConfMix. Among them, $\lambda_{KL} = 0.1$ achieved the best performance, indicating that a lighter regularization is more suitable in this setting. Stronger regularization may restrict adaptation by enforcing excessive similarity between teacher and student predictions.

Figure 11 presents qualitative detection results for selected methods. The examples illustrate the increased difficulty of this scenario, with more frequent missed detections

and lower confidence predictions, particularly in complex urban regions.

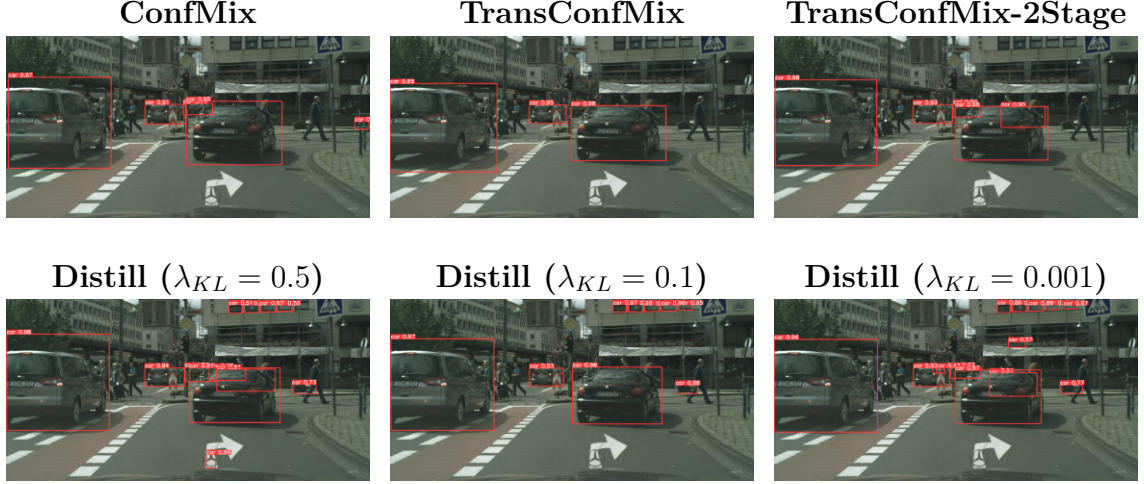


Figure 11 – Qualitative comparison of detection results for the DOLPHINS → Cityscapes scenario. Each panel shows representative predictions obtained with the evaluated adaptation methods on the same target-domain image

6.2.3 SIM10K to nuScenes

Table 6 – Detection performance for the SIM10K → nuScenes adaptation scenario

Method	mAP@0.5	mAP@[0.5:0.95]
Baseline	0.3573	0.1459
CycleGAN	0.3780	0.1505
ConfMix	0.4126	0.1619
TransConfMix	0.3887	0.1570
TransConfMix-2Stage	0.3827	0.1496
ConfMix-Distill ($\lambda_{KL} = 0.5$)	0.3920	0.1565
ConfMix-Distill ($\lambda_{KL} = 0.1$)	0.3917	0.1552
ConfMix-Distill ($\lambda_{KL} = 0.001$)	0.3913	0.1573
Oracle	0.5215	0.3046

Table 6 presents the results for the SIM10K → nuScenes adaptation scenario. In contrast to the SIM10K → Cityscapes case, the performance improvements obtained through DA are more moderate, suggesting that the domain gap between SIM10K and nuScenes is less pronounced or presents different characteristics.

All evaluated adaptation strategies improved over the source-only baseline in mAP@0.5, while most methods also produced consistent, though smaller, gains in mAP@[0.5:0.95]. This indicates that, although adaptation is beneficial, the margin for improvement is more limited in this scenario.

ConfMix achieved the best overall performance, confirming the effectiveness of feature-level adaptation in this domain pair. However, the relative improvement over the baseline

is smaller than in previous scenarios, suggesting that confidence-guided pseudo-labeling is less beneficial when the initial domain discrepancy is less severe or more complex.

CycleGAN-based training provided only modest gains, indicating that pixel-level adaptation alone has a limited impact in this setting. This behavior contrasts with the DOLPHINS $\rightarrow$ Cityscapes scenario, where translation played a more significant role.

The TransConfMix variants achieved competitive results but did not surpass ConfMix. In particular, the single-stage variant slightly outperformed the two-stage version, suggesting that the integration of translated images may have a more limited contribution in this domain pair.

The KL-regularized variants consistently improved over the baseline but remained below the performance of ConfMix. The results show only small variations across different values of λ_{KL} , indicating that the impact of the distillation term is less pronounced in this scenario. This suggests that, when the adaptation task is less challenging, additional regularization provides limited benefit.

Figure 12 presents qualitative detection results for selected methods. The visual comparison highlights that, although most methods produce similar predictions, differences can still be observed in object localization accuracy and confidence, particularly in cluttered scenes.

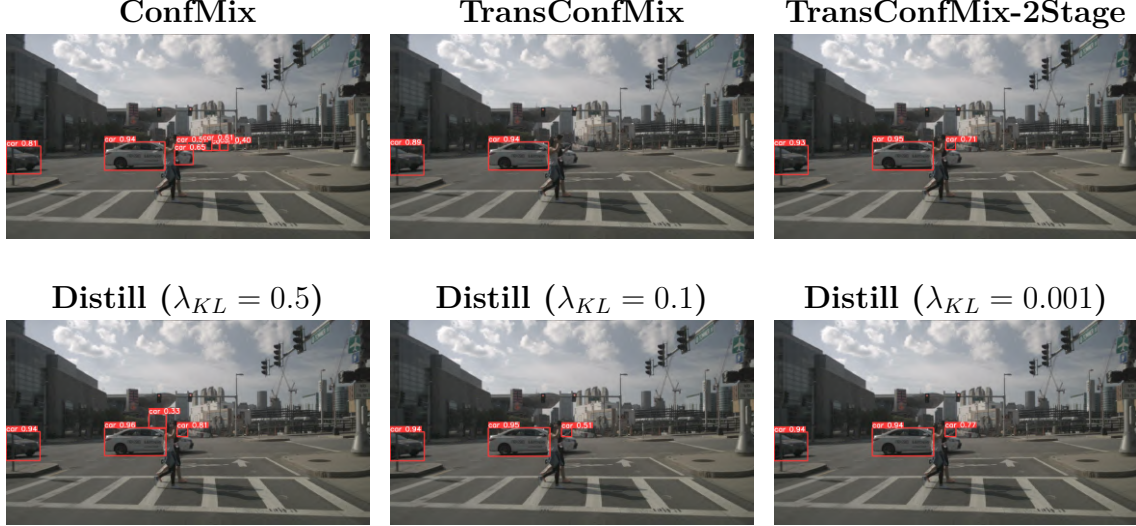


Figure 12 – Qualitative comparison of detection results for the SIM10K $\rightarrow$ nuScenes scenario. Each panel shows representative predictions obtained with the evaluated adaptation methods on the same target-domain image

6.2.4 DOLPHINS to nuScenes

Table 7 presents the results for the DOLPHINS $\rightarrow$ nuScenes adaptation scenario. In contrast to the previous experiments, this setting reveals a markedly different behavior across adaptation strategies.

Table 7 – Detection performance for the DOLPHINS $\rightarrow$ nuScenes adaptation scenario

Method	mAP@0.5	mAP@[0.5:0.95]
Baseline	0.2873	0.1518
CycleGAN	0.2142	0.1087
ConfMix	0.3766	0.1848
TransConfMix	0.3131	0.1379
TransConfMix-2Stage	0.2919	0.1273
ConfMix-Distill ($\lambda_{KL} = 0.5$)	0.3083	0.1381
ConfMix-Distill ($\lambda_{KL} = 0.1$)	0.2991	0.1335
ConfMix-Distill ($\lambda_{KL} = 0.001$)	0.2787	0.1212
Oracle	0.5215	0.3046

ConfMix achieved the best performance in both evaluation metrics, indicating that feature-level adaptation remains robust even when transferring between domains with distinct characteristics. The improvement over the baseline is substantial, highlighting the effectiveness of confidence-guided pseudo-labeling in this scenario.

A notable pattern is observed for CycleGAN-based training, which performs substantially worse than the baseline, suggesting that the unpaired translation process corrupted the structural integrity of the objects, thereby introducing noise into the supervised loss signal. The degradation obtained when using exclusively translated images suggests that the generative model may introduce semantic artifacts, such as local shape distortions or object omissions, that corrupt the supervised loss signal. In this setting, pixel-level translation alone is not sufficient to support robust adaptation and may even impair detector training by weakening the alignment between translated images and the original source annotations.

The TransConfMix variants showed limited effectiveness. Although the single-stage version improved over the baseline in mAP@0.5, both variants failed to consistently improve mAP@[0.5:0.95], indicating reduced localization accuracy. This suggests that combining translated images with feature-level adaptation may not be beneficial when the translation quality is insufficient.

Similarly, the KL-regularized variants did not provide consistent improvements. While some configurations improved mAP@0.5, all of them underperformed relative to ConfMix and most remained below the baseline in mAP@[0.5:0.95]. This behavior indicates that the distillation signal may propagate errors when the teacher predictions are not well aligned with the target domain.

Overall, these results highlight that the effectiveness of pixel-level and hybrid adaptation strategies is highly dependent on the quality of the image translation process. When translation introduces distortions, feature-level methods such as ConfMix become more reliable for DA.

Figure 13 presents qualitative detection results for selected methods. The examples illustrate the challenges of this scenario and show how different methods affect detection

robustness, particularly in complex scenes and under varying visual conditions.



Figure 13 – Qualitative comparison of detection results for the DOLPHINS $\rightarrow$ nuScenes scenario. Each panel shows representative predictions obtained with the evaluated adaptation methods on the same target-domain image

6.3 Summary of Results

Across all evaluated scenarios, the experimental results demonstrate that DA consistently improves detection performance over the source-only baseline. However, the effectiveness of each method varies depending on the characteristics of the source and target domains.

Feature-level adaptation, particularly ConfMix, achieved the most consistent performance across all scenarios, indicating its robustness under different domain gaps. In contrast, pixel-level adaptation using CycleGAN showed mixed results, with improvements in some cases and performance degradation in others.

The proposed variants, including TransConfMix and the KL-regularized models, provided competitive results in certain scenarios but did not consistently outperform the original ConfMix method. These findings suggest that incorporating translated images or additional regularization requires careful consideration of domain characteristics.

Overall, the results highlight that successful DA depends not only on the choice of method but also on the nature of the domain shift, reinforcing the importance of combining complementary adaptation strategies.

Additional evaluation metrics, training time, and more details are provided in Appendix B.

A complementary view emerges when detection performance is decomposed into precision and recall, reported in Appendix B.2 (Tables 12–15). Across scenarios, the source-only baseline tends to behave conservatively, achieving high precision but low recall: in

SIM10K→Cityscapes, for instance, it reaches a precision of 0.854 but a recall of only 0.335, indicating that it detects few vehicles while rarely producing false positives. Domain adaptation consistently shifts this balance, substantially increasing recall at the cost of some precision, which results in higher F1-scores overall. This effect is most pronounced for the KL-regularized variants, which attain the highest recall but also the lowest precision, suggesting that the distillation term leads the detector to accept a larger number of less certain detections.

Notably, the ranking induced by the F1-score does not always coincide with the one induced by mAP. In SIM10K→Cityscapes, for example, TransConfMix-2Stage achieves the highest F1-score (0.529) among the methods for which these metrics are available, even though ConfMix attains the best mAP (the corresponding values for ConfMix are not available, as they derive from the previously published SIBGRAPAPI experiments, where these metrics were not reported). This indicates that the relative merit of each method depends in part on the chosen operating point and metric, reinforcing that no single strategy dominates across all evaluation criteria.

Finally, the training times reported in Appendix B.2 reveal a practical trade-off of pixel-level adaptation. As noted in that appendix, the times listed for ConfMix-based methods correspond only to the adaptation stage and exclude the initial source-only training, so they are not directly comparable to the single-stage baseline. Even so, the cost of training a dedicated CycleGAN model for each domain pair is substantial on its own: in SIM10K→Cityscapes, for example, CycleGAN training alone (43:05) exceeds the entire baseline training (12:05) by more than threefold. Since the benefit of translation was found to be inconsistent across domains, this additional cost becomes a relevant factor when choosing an adaptation strategy in practice.

Chapter 7

Conclusion

This dissertation investigated the Sim2Real object detection problem from the perspective of UDA, with emphasis on the interaction between pixel-level image translation and feature-level adaptation. The study focused on vehicle detection using synthetic source domains and real-world target domains, and evaluated how different adaptation strategies affect detection performance under domain shift. The experimental analysis considered reference models, CycleGAN-based image translation, the original ConfMix framework, the proposed TransConfMix variants, and the ConfMix-Distill regularized extension.

7.1 Summary of Findings

The results obtained in this dissertation show that Sim2Real DA can substantially improve object detection performance when compared with source-only training, although the magnitude of the improvement depends on the source-target pair and on the adaptation strategy employed. Across all evaluated scenarios, the baseline model trained only on synthetic data was consistently outperformed by at least some of the adaptation methods, confirming the practical relevance of DA for reducing the gap between simulated and real-world environments.

The CycleGAN hyperparameter evaluation also showed that image translation quality cannot be treated as a purely visual criterion. Different translation configurations produced different downstream detection results, and the best CycleGAN setting varied across domain pairs. This finding supports the decision to select translation models according to their impact on the final detection task rather than by visual inspection alone.

When the compared adaptation methods are considered, ConfMix produced the strongest overall results among the evaluated adaptation strategies. In all four Sim2Real

scenarios, the original ConfMix framework achieved the best or most consistent performance across both $\text{mAP}@0.5$ and $\text{mAP}@[0.5:0.95]$. This indicates that confidence-guided feature-level adaptation remains a particularly effective strategy for object detection under domain shift.

The proposed TransConfMix methods demonstrated that incorporating translated source images into the adaptation pipeline can be beneficial, but the effect was not uniform across scenarios. TransConfMix and TransConfMix-2Stage generally improved over the source-only baseline and, in some cases, also improved over CycleGAN-only training. However, these gains were not sufficient to consistently surpass the original ConfMix method. A likely explanation is that confidence-guided pseudo-labeling in one-stage detectors benefits from high spatial and structural fidelity in the supervised source branch. Although most translated images preserved the main semantic content of the scene, local degradations introduced by unpaired image translation—especially in small or distant objects—may weaken the correspondence between translated samples and the original source annotations. In this sense, translated synthetic images can help reduce the visual gap between domains, but may also introduce supervision noise that limits the effectiveness of adaptation.

The KL-regularized extension of ConfMix produced mixed results. In some domain pairs, the KL-based variants improved over the source-only baseline and remained competitive with other translated-image strategies. However, the regularization did not consistently outperform the original ConfMix framework, and its effect varied depending on both the target domain and the chosen value of λ_{KL} . These results indicate that consistency regularization between predictions on original and translated source images is a promising idea, but that its effectiveness depends on the training dynamics and on the adaptation scenario.

Taken together, the findings of this dissertation suggest three main conclusions. First, feature-level adaptation was more reliable than pixel-level translation alone in the evaluated Sim2Real settings. Second, translated images can still play a useful supporting role, especially when they improve the visual proximity between source and target domains. Third, although the proposed extensions did not consistently surpass ConfMix, they helped clarify the conditions under which translated data and auxiliary regularization may or may not improve adaptation performance. In this sense, the dissertation contributes not only new adaptation methods, but also a systematic empirical analysis of their behavior across multiple synthetic-to-real domain pairs.

7.2 Research Questions and Hypotheses Revisited

This section revisits the research questions and hypotheses presented in Chapter 1 in light of the experimental findings.

Research Question 1

RQ1: To what extent does the Sim2Real domain shift affect the performance of object detectors trained exclusively on synthetic datasets?

The results clearly show that the Sim2Real domain shift has a strong impact on object detection performance. In all evaluated scenarios, the baseline models trained only on synthetic data remained substantially below the oracle reference, confirming that direct transfer from simulation to real-world images leads to significant performance degradation. This effect was particularly evident in the DOLPHINS $\rightarrow$ Cityscapes scenario, where the baseline achieved the lowest results among all source-only settings.

Central Hypothesis

The performance degradation observed when transferring object detection models from simulation to real-world environments can be significantly reduced by combining high-quality pixel-level domain translation with improved consistency-based feature adaptation.

The results confirm the first part of the central hypothesis, namely that the performance degradation caused by synthetic-to-real transfer is substantial and must be explicitly addressed through adaptation strategies. However, the experiments do not support the stronger claim that this degradation can be significantly reduced by the specific combination proposed in this dissertation in a consistently superior way. Instead, the results show that adaptation is necessary, but the effectiveness of each strategy depends strongly on the interaction between the adaptation mechanism and the characteristics of the domain shift.

Research Question 2

RQ2: How does the quality of pixel-level domain translation influence object detection performance in Sim2Real transfer scenarios?

The results indicate that the impact of pixel-level translation depends strongly on translation quality and on the source-target pair. In some scenarios, such as SIM10K $\rightarrow$ Cityscapes and DOLPHINS $\rightarrow$ Cityscapes, CycleGAN-based training improved over the baseline, suggesting that reducing low-level visual discrepancies can benefit downstream detection. However, this effect was not consistent across all domain pairs. In DOLPHINS $\rightarrow$ nuScenes, CycleGAN-based training performed substantially worse than the baseline, indicating that translated images may introduce distortions or semantic inconsistencies that weaken the supervised training signal.

Hypothesis 1

H1: Systematic optimization of CycleGAN hyperparameters and architectural configurations improves the semantic consistency between translated synthetic images and real target samples, which in turn benefits downstream object detection performance.

Therefore, H1 is only partially supported. The experiments do indicate that the selection and optimization of CycleGAN configurations matter for downstream performance, since different translation settings produced different detection results and the best configuration varied across domain pairs. However, the results do not support a uniformly positive relationship between translation quality and object detection gains. Instead, they suggest that translation is useful only when it preserves the structural information required for detection.

Research Question 3

RQ3: Can consistency-based DA frameworks be improved by introducing additional regularization mechanisms, such as KL-divergence constraints, to stabilize pseudo-label learning?

The results do not provide evidence that the proposed KL-based regularization consistently improves the original ConfMix framework. Although the ConfMix-Distill variants often remained competitive and, in some cases, improved over the baseline, they did not surpass ConfMix in any of the evaluated scenarios. This suggests that the proposed regularization, in its current form, was not sufficient to produce systematic gains.

Hypothesis 2

H2: Incorporating an additional Kullback–Leibler (KL) divergence regularization term into a consistency-based adaptation framework improves prediction calibration and reduces the propagation of errors during pseudo-label refinement.

Accordingly, H2 is not confirmed by the present experiments. While the addition of a KL term is a valid and technically feasible extension, its current formulation did not consistently improve prediction stability or adaptation performance. A plausible interpretation is that constraining only the objectness predictions limits the effect of the regularization on the final detection behavior, which also depends on classification and localization quality. In addition, the teacher–student formulation may propagate source-domain bias when the teacher itself is trained only on synthetic data.

Research Question 4

RQ4: How do pixel-level domain translation and feature-level adaptation strategies interact when applied to modern object detection architectures in Sim2Real settings?

The interaction between pixel-level translation and feature-level adaptation proved to be scenario-dependent. The experiments show that feature-level adaptation, represented by ConfMix, was the most robust strategy across all four scenarios. In contrast, the hybrid variants that combine translated images with confidence-guided adaptation produced mixed results. In some cases they improved over the baseline and over CycleGAN-only training, but they did not consistently exceed the performance of ConfMix.

Hypothesis 3

H3: Pixel-level alignment and feature-level consistency regularization provide complementary advantages, and their joint application leads to better Sim2Real transfer performance than either strategy used independently.

These findings do not support H3 in its original form. The expected complementarity between pixel-level alignment and additional consistency-based regularization was not consistently observed. Instead, the results show that the combination is highly sensitive to translation fidelity and domain characteristics. When the translated images preserve relevant scene structure, they may support adaptation; when they introduce distortions, they may reduce the reliability of the supervision and weaken the benefits of confidence-guided learning.

Integrated Interpretation

Taken together, the results suggest that the original ConfMix framework offers the most reliable balance between adaptation strength and training stability in the evaluated Sim2Real settings. A likely reason is that ConfMix operates directly on the adaptation problem through confidence-guided pseudo-labeling, without depending on an intermediate image translation step that may alter object geometry or local visual structure. In contrast, the proposed TransConfMix variants rely on the assumption that translated images remain sufficiently aligned with the original source annotations. When this assumption is weakened by artifacts, object omissions, or local distortions, the resulting supervision becomes less reliable.

This interpretation is consistent with the poorer results observed in scenarios where translated images seem to be less effective, particularly DOLPHINS $\rightarrow$ nuScenes. In this case, the stronger performance of ConfMix suggests that preserving the original synthetic supervision and relying on confidence-guided feature-level adaptation is more robust than introducing translated source samples with uncertain structural fidelity.

Overall, the results indicate that the main contribution of this dissertation is not the demonstration of a universally superior hybrid adaptation method, but rather a systematic empirical analysis of when pixel-level translation may support or hinder confidence-based UDA for one-stage object detection in Sim2Real settings.

7.3 Limitations

Although the results obtained in this dissertation provide relevant evidence about Sim2Real DA for object detection, this study has limitations that should be acknowledged.

First, the experiments were restricted to vehicle detection. Although this choice ensures label compatibility across all source and target domains and allows controlled comparison between methods, it also limits the generality of the conclusions. Restricting the adaptation problem to a single class simplifies the learning setting, since the model is effectively required to distinguish mainly between foreground and background. In more complex multi-class scenarios, the behavior of the evaluated strategies may differ substantially, especially for categories with greater intra-class variation or large scale changes, such as pedestrians and traffic signs. In such cases, both pixel-level translation and confidence-guided pseudo-labeling may become more sensitive to local distortions, small-object degradation, and class-specific ambiguities, which were not fully captured in the present study.

Second, the experimental evaluation was conducted on a limited number of synthetic-to-real domain pairs. While the selected scenarios include meaningful variations in both source and target distributions, they do not cover the full diversity of possible Sim2Real conditions. As a result, the conclusions drawn here should be interpreted within the scope of the evaluated datasets and not as universal statements about all object detection adaptation problems.

Third, the study adopted YOLOv5 as the single detection architecture in all experiments, mainly to preserve reproducibility and to ensure a fair comparison with the original ConfMix framework. This choice was important for controlled evaluation, since it avoids confounding the effect of the proposed adaptation strategies with differences introduced by changing the underlying detector architecture. However, YOLOv5 is no longer among the most recent one-stage detectors, and the observed behavior of the proposed methods may depend in part on architectural characteristics specific to this model family. Therefore, different results may be obtained when applying the same ideas to more recent detector architectures or to other detector families with different design choices.

Another limitation concerns the selection and use of translated images. Although the CycleGAN configurations were chosen according to downstream detection performance rather than visual inspection alone, the translation process remains sensitive to hyperparameter choices and to the characteristics of each domain pair. In addition, the use of translated synthetic images does not guarantee semantic or geometric preservation in all cases. Unpaired image translation may introduce local distortions, hallucinations, or object omissions, which can weaken the correspondence between translated images and the original source annotations reused during training. Since this dissertation did not include a dedicated post-translation geometric consistency analysis, the potential mismatch between translated samples and their original bounding-box labels should be regarded as

an important limitation of the study.

Another limitation concerns the stability of the CycleGAN selection protocol with respect to the subset of target-domain images used during evaluation. Since the translated-image configurations were compared through downstream detection performance on a reduced sample of the target domain, the selected model may depend to some extent on the specific subset considered. In principle, a different sample could favor a different translation configuration, especially when the target domain is visually diverse. Therefore, although the adopted protocol is task-driven and practically useful, the selected CycleGAN configuration should be interpreted as the best one within the evaluated sampling setting rather than as a universally optimal choice for the entire target distribution.

A related methodological point is that the CycleGAN configurations were selected according to downstream detector performance measured on target-domain data. Although this protocol does not use target annotations for training the adaptation methods themselves, it does make the model selection process dependent on the evaluation domain. For this reason, one may argue that the procedure does not preserve a strict separation between translation-model selection and final target-domain evaluation. In this dissertation, the protocol was adopted because it aligns translation selection with the final detection objective; however, this choice should still be acknowledged as a potential source of selection bias in the experimental design.

The KL-regularized extension also presents limitations. In the evaluated scenarios, its effect was not consistent across domain pairs or across different values of λ_{KL} . This indicates that the proposed regularization strategy still depends strongly on training conditions and may require further calibration to produce more stable gains. In addition, the current formulation applies the distillation term only to the objectness predictions, without directly regularizing class probabilities, bounding-box regression outputs, or intermediate feature representations. Since detection confidence in one-stage detectors depends not only on objectness but also on classification and localization quality, restricting the regularization to objectness alone may limit the influence of the method on the final detection behavior. Therefore, although the method is conceptually promising, its current formulation should be interpreted as an exploratory extension rather than as a fully validated improvement over the original ConfMix framework.

Finally, part of the experimental study depends on empirical choices such as the CycleGAN search space, the selected regularization weights, and the training schedules adopted for the adaptation methods. While these choices were made systematically and under controlled conditions, they still define the boundaries of the conclusions reported in this dissertation.

7.4 Future Work

Several directions can be explored to extend the research presented in this dissertation.

A first natural direction is to evaluate the proposed methods in broader object detection settings, including multi-class scenarios and additional source-target domain pairs. This would help determine whether the observed behavior of ConfMix, TransConfMix, and KL-based regularization remains consistent when the task becomes more complex and the semantic gap between classes increases.

Another relevant direction is to investigate the proposed strategies with other detector architectures. Since all experiments in this dissertation were conducted with YOLOv5 in order to maintain consistency with the original ConfMix framework, it would be valuable to assess whether the same conclusions hold for more recent one-stage detectors or for architectures with different design choices.

The use of translated images can also be explored more deeply. In the present work, translated synthetic samples were incorporated into the adaptation pipeline in predefined ways, but other integration strategies may be investigated. For example, future work may consider adaptive policies for selecting when and how translated images should be introduced during training, as well as alternative forms of combining original and translated source samples.

A related direction concerns the image translation model itself. Although CycleGAN was adopted in this dissertation due to its suitability for unpaired translation, other generative models may offer improved semantic preservation or better stability during training. Comparing different image translation frameworks could provide further insight into the role of pixel-level adaptation in Sim2Real transfer.

A further direction for future investigation is to incorporate explicit image-quality metrics such as the Fréchet Inception Distance (FID) into translation model evaluation. In the present work, CycleGAN configurations were selected according to downstream detection performance, which aligns model selection with the final task but is also computationally expensive. Future studies may investigate whether improvements in visual-domain similarity, as measured by FID or related metrics, correlate with gains in object detection performance. Such an analysis could help clarify to what extent image-level realism reflects detection-oriented utility in Sim2Real adaptation.

Another promising direction is to incorporate detector-guided consistency directly into the training of the image translation model. One possible strategy is to first train a detector on the original source domain and then freeze this detector while optimizing the translation model. An additional consistency term could then be imposed between the detector outputs obtained from the original source images and from their translated counterparts. Such a formulation could encourage the translation model not only to transfer visual style, but also to preserve detection-relevant semantic and geometric structure, which may reduce the mismatch between translated images and the original

source annotations reused during training.

The KL-based extension also opens several possibilities for future investigation. One direction is to study alternative forms of consistency regularization beyond objectness, including losses applied to class predictions, bounding-box regression outputs, or intermediate feature representations. This extension is particularly relevant because image translation artifacts may affect not only object confidence, but also object boundaries and localization quality. Another important direction is to investigate more robust teacher-student update strategies in order to reduce the propagation of domain bias from the teacher to the student. In particular, future work may explore teacher models updated through exponential moving average mechanisms, rather than relying only on direct prediction distillation from a fixed teacher. It would also be valuable to examine dynamic weighting schedules for the KL term, such as warm-up strategies in which its contribution starts near zero and increases as adaptation becomes more stable. Such mechanisms may help reduce confirmation bias and improve the effectiveness of regularization during pseudo-label-based training. Beyond these mechanisms, a more systematic and fine-grained search over the regularization weight λ_{KL} — for example, using a denser logarithmic grid than the limited set of representative values evaluated in this dissertation — could further clarify whether more stable and consistent gains can be obtained from the distillation term across the different domain pairs.

Future work may also investigate whether translated-image supervision and feature-level adaptation can be coupled more tightly, rather than being combined through relatively simple training-stage modifications. More integrated formulations may allow translated samples to contribute more effectively to the adaptation process.

Finally, a promising long-term direction is to evaluate these methods in more realistic deployment conditions, including scenarios with stronger environmental variability, additional sensing conditions, or restrictions related to real-world annotation availability. Such studies would help clarify the practical potential of Sim2Real DA for traffic object detection in real applications.

Part of the contributions of this dissertation has been published in the proceedings of SIBGRAPI by Medeiros et al. (2025), one of the main conferences in computer vision and pattern recognition in Latin America. The published work presents the proposed TransConfMix, along with their evaluation in Sim2Real object detection scenarios. This publication provides independent validation of the proposed methods and highlights their relevance to the research community.

References

- ARRUDA, V. F. et al. Cross-domain object detection using unsupervised image translation. **Expert Systems with Applications**, Elsevier, v. 192, p. 116334, 2022.
- BEN-DAVID, S. et al. A theory of learning from different domains. **Machine Learning**, v. 79, n. 1, p. 151–175, 2010.
- CAESAR, H. et al. nusenes: A multimodal dataset for autonomous driving. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Seattle, WA, USA: IEEE, 2020.
- CHEN, L.; HAN, J.; WANG, Y. Datr: Unsupervised domain adaptive detection transformer with dataset-level adaptation and prototypical alignment. **IEEE Transactions on Image Processing**, v. 34, p. 982–994, 2025.
- CHEN, Y. et al. Domain adaptive faster r-cnn for object detection in the wild. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Salt Lake City, UT, USA: IEEE, 2018.
- CORDTS, M. et al. The cityscapes dataset for semantic urban scene understanding. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Las Vegas, NV, USA: IEEE, 2016.
- CSURKA, G. A comprehensive survey on domain adaptation for visual applications. In: **Domain Adaptation in Computer Vision Applications**. Cham: Springer, 2017, (Advances in Computer Vision and Pattern Recognition). p. 1–35.
- GANIN, Y.; LEMPITSKY, V. Unsupervised domain adaptation by backpropagation. In: **Proceedings of the International Conference on Machine Learning (ICML)**. Lille, France: PMLR, 2015.
- GANIN, Y. et al. Domain-adversarial training of neural networks. **Journal of Machine Learning Research**, v. 17, n. 59, p. 1–35, 2016.
- GIRSHICK, R. et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Columbus, OH, USA: IEEE, 2014.
- HOFFMAN, J. et al. Cycada: Cycle-consistent adversarial domain adaptation. In: **Proceedings of the International Conference on Machine Learning (ICML)**. Stockholm, Sweden: PMLR, 2018.

- ISOLA, P. et al. Image-to-image translation with conditional adversarial networks. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Honolulu, HI, USA: IEEE, 2017.
- JOHNSON-ROBERSON, M. et al. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In: **Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)**. Stockholm, Sweden: IEEE, 2016.
- LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 2015.
- LIU, L. et al. Deep learning for generic object detection: A survey. **International Journal of Computer Vision**, v. 128, n. 2, p. 261–318, 2020.
- LIU, W. et al. Ssd: Single shot multibox detector. In: **Proceedings of the European Conference on Computer Vision (ECCV)**. Amsterdam, Netherlands: Springer, 2016.
- LONG, M. et al. Learning transferable features with deep adaptation networks. In: **Proceedings of the International Conference on Machine Learning (ICML)**. Lille, France: PMLR, 2015.
- Long, Mingsheng et al. Conditional adversarial domain adaptation. **Advances in Neural Information Processing Systems**, Curran Associates, Inc., Montreal, Canada, v. 31, 2018.
- MAO, R. et al. Dolphins: Dataset for collaborative perception enabled harmonious and interconnected self-driving. In: **Proceedings of the Asian Conference on Computer Vision (ACCV)**. Macao, China: Springer, 2022. p. 495–511.
- MARCUS, R. et al. Synth it like kitti: Synthetic data generation for object detection in driving scenarios. In: **Proceedings of the International Conference on Robotics, Computer Vision and Intelligent Systems (ROBOVIS)**. Porto, Portugal: Springer, 2025. p. 414–432.
- MATTOLIN, G. et al. Confmix: Unsupervised domain adaptation for object detection via confidence-based mixing. In: **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**. Waikoloa, HI, USA: IEEE, 2023. p. 423–433.
- MEDEIROS, A. N. et al. Adapting synthetic eyes: A study of pixel and feature-level uda for traffic object detection. In: **Proceedings of the SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. Salvador, BA, Brazil: IEEE, 2025. p. 1–6.
- PENG, X. et al. Syn2real: A new benchmark for synthetic-to-real visual domain adaptation. **arXiv preprint arXiv:1806.09755**, 2018.
- PENG, X. B. et al. Sim-to-real transfer of robotic control with dynamics randomization. In: **Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)**. Brisbane, Australia: IEEE, 2018.

- REDMON, J. et al. You only look once: Unified, real-time object detection. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Las Vegas, NV, USA: IEEE, 2016.
- REDMON, J.; FARHADI, A. Yolov3: An incremental improvement. **CoRR**, abs/1804.02767, 2018. Disponível em: <<http://arxiv.org/abs/1804.02767>>.
- REN, S. et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In: **Proceedings of the 29th International Conference on Neural Information Processing Systems**. Montreal, Canada: MIT Press, 2015.
- RICHTER, S. et al. Playing for data: Ground truth from computer games. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Las Vegas, NV, USA: IEEE, 2016.
- ROS, G. et al. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Las Vegas, NV, USA: IEEE, 2016.
- SADEGHIGOL, B.; KEYVANRAD, M. A. Bridging the synthetic-to-real gap (bsrg): Creating simulated datasets for domain adaptation to enhance vehicle detection. In: **Proceedings of the International Conference on Computer and Knowledge Engineering (ICCKE)**. Mashhad, Iran: IEEE, 2024.
- SAITO, K. et al. Strong-weak distribution alignment for adaptive object detection. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Long Beach, CA, USA: IEEE, 2019.
- Saito, Kuniaki et al. Maximum classifier discrepancy for unsupervised domain adaptation. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Salt Lake City, UT, USA: IEEE, 2018.
- SAPKOTA, R.; KARKEE, M. Ultralytics yolo evolution: An overview of yolo26, yolo11, yolov8 and yolov5 object detectors for computer vision and pattern recognition. **arXiv preprint arXiv:2510.09653**, abs/2510.09653, 2025.
- SULTANA, F.; SUFIAN, A.; DUTTA, P. A review of object detection models based on convolutional neural network. In: **Intelligent Computing: Image Processing Based Applications**. Singapore: Springer, 2019. p. 1–15.
- SUN, B.; SAENKO, K. Deep coral: Correlation alignment for deep domain adaptation. In: **Proceedings of the European Conference on Computer Vision (ECCV)**. Amsterdam, Netherlands: Springer, 2016. p. 443–450.
- TOBIN, J. et al. Domain randomization for transferring deep neural networks from simulation to the real world. In: **Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)**. Vancouver, BC, Canada: IEEE, 2017.
- TORRALBA, A.; EFROS, A. A. Unbiased look at dataset bias. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Colorado Springs, CO, USA: IEEE, 2011. p. 1521–1528.

- TREMBLAY, J. et al. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Salt Lake City, UT, USA: IEEE, 2018.
- TSAI, Y.-H. et al. Learning to adapt structured output space for semantic segmentation. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Salt Lake City, UT, USA: IEEE, 2018.
- TZENG, E. et al. Adversarial discriminative domain adaptation. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Honolulu, HI, USA: IEEE, 2017.
- VAROL, G. et al. Learning from synthetic humans. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Honolulu, HI, USA: IEEE, 2017. p. 4627–4635.
- WANG, C.-Y.; BOCHKOVSKIY, A.; LIAO, H.-Y. M. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. Vancouver, BC, Canada: IEEE, 2023. p. 7464–7475.
- WANG, J. et al. A parallel teacher for synthetic-to-real domain adaptation of traffic object detection. **IEEE Transactions on Intelligent Vehicles**, v. 7, n. 3, p. 441–455, 2022.
- WANG, M.; DENG, W. Deep visual domain adaptation: A survey. **Neurocomputing**, v. 312, p. 135–153, 2018.
- XIE, Q. et al. Self-training with noisy student improves imagenet classification. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Seattle, WA, USA: IEEE, 2020.
- YU, F. et al. Sc-uda: Style and content gaps aware unsupervised domain adaptation for object detection. In: **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**. Waikoloa, HI, USA: IEEE, 2022. p. 1061–1070.
- YU, X.; LU, X. Domain adaptation of anchor-free object detection for urban traffic. **Neurocomputing**, v. 582, p. 127477, 2024.
- YUN, S. et al. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**. Seoul, South Korea: IEEE, 2019.
- ZHU, J.-Y. et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: **Proceedings of the IEEE International Conference on Computer Vision (ICCV)**. Venice, Italy: IEEE, 2017.
- ZOU, Y. et al. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: **Proceedings of the European Conference on Computer Vision (ECCV)**. Munich, Germany: Springer, 2018.

ZOU, Z. et al. Object detection in 20 years: A survey. **Proceedings of the IEEE**, v. 111, n. 3, p. 257–276, 2023.

Appendices

Appendix A

Detailed CycleGAN Hyperparameter Search

A.1 Evaluated Configurations

Four CycleGAN configurations were evaluated for each source-target adaptation pair in order to determine which image translation settings were most effective for downstream object detection.

The evaluated configurations differed in terms of image preprocessing, learning-rate scheduling, and the use of the identity-loss term.

To make the model selection procedure computationally feasible, the CycleGAN models were initially trained using a subset of the available training data. Specifically, a representative sample corresponding to approximately 10% of each source and target dataset was used during this stage.

After training, each CycleGAN model was applied to translate the entire source dataset to the target domain. A YOLOv5 detector was then trained on the translated images using the original source annotations, and evaluated on the target domain. The resulting detection performance was used as the criterion for selecting the best CycleGAN configuration. After this selection stage, the chosen configuration was retrained using the full source and target datasets. The resulting CycleGAN models were then used to generate the translated datasets employed in the experiments reported in Chapter 6.

□ **Configuration 1:** default preprocessing with image cropping, `load_size=1024`, `crop_size=512`, constant learning-rate schedule, and `lambda_identity=1.0`.

- ❑ **Configuration 2:** no explicit cropping or loading resize parameters, constant learning-rate schedule, and `lambda_identity=1.0`.
- ❑ **Configuration 3:** same image preprocessing settings as Configuration 1, but replacing the default learning-rate schedule with the `plateau` policy.
- ❑ **Configuration 4:** no explicit cropping or loading resize parameters, `plateau` learning-rate schedule, and no identity-loss weighting term.

All configurations were trained with identical remaining settings, including 100 epochs of initial training followed by 100 epochs of learning-rate decay, batch size of 1, learning rate of 0.0001, and no dropout. Since four configurations were evaluated for each of the four adaptation pairs, this stage resulted in a total of 16 translation-and-evaluation runs.

A.2 Complete Results by Domain Pair

Tables 8–11 present the complete detection results obtained when training YOLOv5 detectors on synthetic images translated by different CycleGAN configurations.

Figures 15–18 illustrate representative examples of the translated images for each adaptation pair.

For visual comparison, all translation results are generated from the same reference source images for each dataset. Figure 14 shows the original input samples used for the SIM10K and DOLPHINS datasets.



SIM10K source image



DOLPHINS source image

Figure 14 – Example source images from the SIM10K and DOLPHINS datasets, shown for qualitative reference. These samples illustrate the type of input data used

A.2.1 SIM10K to Cityscapes

Table 8 presents the detection performance obtained with each CycleGAN configuration for this adaptation pair, along with additional evaluation metrics and computational cost. Precision, recall, and F1-score are computed at an IoU threshold of 0.5. The generator time (Gen. Time) corresponds to the CycleGAN training time, while the detector time

(Det. Time) refers to the YOLOv5 training time on the translated dataset. Figure 15 shows representative translation results.

Table 8 – Detection performance, additional metrics, and computational cost obtained using CycleGAN-translated SIM10K images evaluated on Cityscapes

Configuration	mAP@0.5	mAP@[0.5:0.95]	Precision	Recall	F1-score	Gen. Time (hh:mm)	Det. Time (hh:mm)
Configuration 1	0.361157	0.216198	0.763096	0.373470	0.501499	13:56	12:21
Configuration 2	0.372138	0.221599	0.793772	0.377800	0.511939	14:08	11:51
Configuration 3	0.381407	0.230084	0.759987	0.395482	0.520241	15:36	10:32
Configuration 4	0.371276	0.224559	0.778782	0.380415	0.511148	10:03	12:22



Configuration 1



Configuration 2



Configuration 3



Configuration 4

Figure 15 – CycleGAN translation examples for the SIM10K $\rightarrow$ Cityscapes adaptation pair

A.2.2 DOLPHINS to Cityscapes

Table 9 summarizes the detection performance, additional metrics, and computational cost obtained with each CycleGAN configuration for this adaptation pair. Precision, recall, and F1-score are computed at an IoU threshold of 0.5. The generator time (Gen. Time) corresponds to the CycleGAN training time, while the detector time (Det. Time) refers to the YOLOv5 training time on the translated dataset. Figure 16 shows representative translation results.

Table 9 – Detection performance, additional metrics, and computational cost obtained using CycleGAN-translated DOLPHINS images evaluated on Cityscapes

Configuration	mAP@0.5	mAP@[0.5:0.95]	Precision	Recall	F1-score	Gen. Time (hh:mm)	Det. Time (hh:mm)
Configuration 1	0.169828	0.086397	0.812656	0.172256	0.284258	9:48	9:51
Configuration 2	0.145143	0.076187	0.885478	0.144120	0.247893	2:48	9:15
Configuration 3	0.159535	0.078417	0.803493	0.164639	0.273281	9:30	9:15
Configuration 4	0.150777	0.077364	0.800759	0.151717	0.255100	1:54	9:53



Configuration 1



Configuration 2



Configuration 3



Configuration 4

Figure 16 – CycleGAN translation examples for the DOLPHINS $\rightarrow$ Cityscapes adaptation pair

A.2.3 SIM10K to nuScenes

Table 10 presents the detection performance, additional metrics, and computational cost obtained with each CycleGAN configuration for this adaptation pair. Precision, recall, and F1-score are computed at an IoU threshold of 0.5. The generator time (Gen. Time) corresponds to the CycleGAN training time, while the detector time (Det. Time) refers to the YOLOv5 training time on the translated dataset. Figure 17 shows representative translation results.

Table 10 – Detection performance, additional metrics, and computational cost obtained using CycleGAN-translated SIM10K images evaluated on nuScenes

Configuration	mAP@0.5	mAP@[0.5:0.95]	Precision	Recall	F1-score	Gen. Time (hh:mm)	Det. Time (hh:mm)
Configuration 1	0.363343	0.148039	0.469157	0.441768	0.455051	41:00	12:55
Configuration 2	0.375606	0.145592	0.478608	0.453353	0.465638	9:30	12:40
Configuration 3	0.378564	0.153247	0.462762	0.453787	0.458231	44:00	12:56
Configuration 4	0.383473	0.154258	0.498521	0.456566	0.476622	16:00	12:44



Configuration 1



Configuration 2



Configuration 3



Configuration 4

Figure 17 – CycleGAN translation examples for the SIM10K $\rightarrow$ nuScenes adaptation pair

A.2.4 DOLPHINS to nuScenes

Table 11 presents the detection performance, additional metrics, and computational cost obtained with each CycleGAN configuration for this adaptation pair. Precision, recall, and F1-score are computed at an IoU threshold of 0.5. The generator time (Gen. Time) corresponds to the CycleGAN training time, while the detector time (Det. Time) refers to the YOLOv5 training time on the translated dataset. Figure 18 shows representative translation results.

Table 11 – Detection performance, additional metrics, and computational cost obtained using CycleGAN-translated DOLPHINS images evaluated on nuScenes

Configuration	mAP@0.5	mAP@[0.5:0.95]	Precision	Recall	F1-score	Gen. Time (hh:mm)	Det. Time (hh:mm)
Configuration 1	0.262817	0.133427	0.671526	0.297612	0.412438	42:00	9:06
Configuration 2	0.213262	0.104719	0.742408	0.232516	0.354124	13:30	10:33
Configuration 3	0.270817	0.136088	0.688331	0.299739	0.417622	41:30	10:21
Configuration 4	0.263516	0.138767	0.675049	0.297606	0.413094	13:44	10:09



Configuration 1



Configuration 2



Configuration 3



Configuration 4

Figure 18 – CycleGAN translation examples for the DOLPHINS $\rightarrow$ nuScenes adaptation pair

Across the evaluated domain pairs, the best-performing CycleGAN configuration varied depending on the source–target combination. For SIM10K $\rightarrow$ Cityscapes, Configuration 3 achieved the best results in both evaluation metrics. For DOLPHINS $\rightarrow$ Cityscapes, the best performance was obtained with Configuration 1. For SIM10K $\rightarrow$ nuScenes, Configuration 4 produced the highest scores in both $\text{mAP}@0.5$ and $\text{mAP}@[0.5:0.95]$. In the DOLPHINS $\rightarrow$ nuScenes scenario, Configuration 3 achieved the highest $\text{mAP}@0.5$, while Configuration 4 obtained the highest $\text{mAP}@[0.5:0.95]$.

These results support the CycleGAN configuration choices adopted in the main experimental pipeline described in Chapter 5.

Appendix B

Additional Results and Implementation Details

B.1 Hardware and Computational Infrastructure

All experiments were conducted using three different computational servers within the laboratory infrastructure, identified as *gpu1*, *ws2*, and *ss2*.

The majority of the training procedures were performed on the *gpu1* server, which served as the primary computational resource. The *ws2* server was used as a secondary resource for additional experiments, while the *ss2* server was used less frequently, mainly for validation and auxiliary tasks.

B.1.1 *gpu1* (Primary Server)

- ❑ **GPU:** $4 \times$ NVIDIA RTX 5000 Ada Generation (32 GB VRAM each)
- ❑ **CPU:** AMD Ryzen Threadripper PRO 5975WX (32 cores, 64 threads)
- ❑ **RAM:** 251 GB
- ❑ **Operating System:** Ubuntu 24.04.2 LTS
- ❑ **CUDA Version:** 12.8

B.1.2 *ws2* (Secondary Server)

- ❑ **GPU:** NVIDIA GeForce RTX 3080 (10 GB VRAM)

- ❑ **CPU:** Intel Xeon E5-2630 v4 (2 sockets, 20 cores total, 40 threads)
- ❑ **RAM:** 125 GB
- ❑ **Operating System:** Ubuntu 22.04.4 LTS
- ❑ **CUDA Version:** 12.5

B.1.3 ss2 (Auxiliary Server)

- ❑ **GPU:** NVIDIA Tesla P100 (16 GB VRAM)
- ❑ **CPU:** Intel Xeon E5-2683 v4 (2 sockets, 32 cores total, 64 threads)
- ❑ **RAM:** 251 GB
- ❑ **Operating System:** Ubuntu 22.04.4 LTS
- ❑ **CUDA Version:** 12.5

B.2 Additional Quantitative Results

This section presents complementary evaluation metrics for all domain adaptation scenarios considered in this dissertation. In addition to mAP, the tables report precision, recall, and F1-score computed at an IoU threshold of 0.5, as well as total training time for each method.

For methods based on ConfMix, the reported training time corresponds only to the domain adaptation stage (50 epochs), and does not include the initial source-only training phase (Step 1), which is also performed for 50 epochs. Therefore, the total training time of ConfMix-based methods is higher than the values reported in the tables.

Additionally, some metrics for ConfMix are not reported in the tables because the results were obtained from previously published experiments in our SIBGRAPI paper (MEDEIROS et al., 2025), where precision, recall, and F1-score were not included. For consistency, the original reported values are preserved.

B.2.1 SIM10K to Cityscapes

Table 12 reports additional evaluation metrics for this scenario, providing a complementary view of detection performance in terms of precision–recall balance and training cost.

Table 12 – Additional evaluation metrics for the SIM10K $\rightarrow$ Cityscapes scenario. Precision, recall, and F1-score are computed at IoU=0.5

Method	mAP (0.5 / 0.5:0.95)	Precision	Recall	F1-score	Time (hh:mm)
Baseline	0.317213 / 0.199802	0.853772	0.334821	0.481007	12:05
CycleGAN	0.362995 / 0.219619	0.769546	0.374481	0.503800	43:05
ConfMix	0.4082 / 0.2396	–	–	–	–
TransConfMix	0.325804 / 0.192418	0.658885	0.388251	0.488595	14:33
TransConfMix-2Stage	0.390438 / 0.228640	0.754728	0.406975	0.528802	14:15
ConfMix-Distill ($\lambda_{KL} = 0.5$)	0.395461 / 0.218069	0.513959	0.429103	0.467713	17:05
ConfMix-Distill ($\lambda_{KL} = 0.1$)	0.369817 / 0.203798	0.521998	0.397991	0.451637	16:00
ConfMix-Distill ($\lambda_{KL} = 0.001$)	0.370371 / 0.212437	0.524855	0.410497	0.460685	16:39

B.2.2 DOLPHINS to Cityscapes

Table 13 reports additional evaluation metrics for this scenario, providing a complementary view of detection performance in terms of precision–recall balance and training cost.

Table 13 – Additional evaluation metrics for the DOLPHINS $\rightarrow$ Cityscapes scenario. Precision, recall, and F1-score are computed at IoU=0.5

Method	mAP (0.5 / 0.5:0.95)	Precision	Recall	F1-score	Time (hh:mm)
Baseline	0.105331 / 0.048118	0.951613	0.088989	0.162759	9:48
CycleGAN	0.164198 / 0.092732	0.881340	0.166456	0.280024	34:20
ConfMix	0.2883 / 0.1413	–	–	–	–
TransConfMix	0.259066 / 0.132862	0.690603	0.275585	0.393961	14:22
TransConfMix-2Stage	0.249383 / 0.118528	0.639115	0.271022	0.380633	14:20
ConfMix-Distill ($\lambda_{KL} = 0.5$)	0.242826 / 0.118088	0.489995	0.301035	0.372946	17:13
ConfMix-Distill ($\lambda_{KL} = 0.1$)	0.271705 / 0.135536	0.478632	0.307991	0.374803	16:00
ConfMix-Distill ($\lambda_{KL} = 0.001$)	0.269337 / 0.124252	0.467223	0.317563	0.378123	17:00

B.2.3 SIM10K to nuScenes

Table 14 reports additional evaluation metrics for this scenario, providing a complementary view of detection performance in terms of precision–recall balance and training cost.

Table 14 – Additional evaluation metrics for the SIM10K $\rightarrow$ nuScenes scenario. Precision, recall, and F1-score are computed at IoU=0.5

Method	mAP (0.5 / 0.5:0.95)	Precision	Recall	F1-score	Time (hh:mm)
Baseline	0.357302 / 0.145901	0.514691	0.419969	0.462530	12:05
CycleGAN	0.378000 / 0.150529	0.495173	0.453416	0.473375	44:06
ConfMix	0.4126 / 0.1619	–	–	–	–
TransConfMix	0.388654 / 0.157000	0.515275	0.473243	0.493365	47:50
TransConfMix-2Stage	0.382746 / 0.149590	0.497821	0.475193	0.486244	48:50
ConfMix-Distill ($\lambda_{KL} = 0.5$)	0.392028 / 0.156508	0.441742	0.476277	0.458360	55:32
ConfMix-Distill ($\lambda_{KL} = 0.1$)	0.391703 / 0.155191	0.427769	0.474473	0.449912	57:00
ConfMix-Distill ($\lambda_{KL} = 0.001$)	0.391277 / 0.157295	0.467393	0.472401	0.469883	57:00

B.2.4 DOLPHINS to nuScenes

Table 15 reports additional evaluation metrics for this scenario, providing a complementary view of detection performance in terms of precision–recall balance and training

cost.

Table 15 – Additional evaluation metrics for the DOLPHINS $\rightarrow$ nuScenes scenario. Precision, recall, and F1-score are computed at IoU=0.5

Method	mAP (0.5 / 0.5:0.95)	Precision	Recall	F1-score	Time (hh:mm)
Baseline	0.287302 / 0.151843	0.739835	0.255681	0.380027	09:48
CycleGAN	0.214236 / 0.108739	0.663093	0.246460	0.359355	36:00
ConfMix	0.3766 / 0.1848	–	–	–	–
TransConfMix	0.313059 / 0.137864	0.584713	0.367634	0.451433	36:56
TransConfMix-2Stage	0.291927 / 0.127256	0.596554	0.339731	0.432919	37:23
ConfMix-Distill ($\lambda_{KL} = 0.5$)	0.308297 / 0.138098	0.454461	0.379816	0.413799	41:32
ConfMix-Distill ($\lambda_{KL} = 0.1$)	0.299142 / 0.133474	0.458816	0.359465	0.403109	40:13
ConfMix-Distill ($\lambda_{KL} = 0.001$)	0.278718 / 0.121152	0.492889	0.346308	0.406797	43:27

Folha de Aprovação

Defesa de dissertação de mestrado do(a) candidato(a) André Neves Medeiros, realizada em 23/04/2026

Comissão Julgadora

Prof(a) Dr(a) Tiago Agostinho de Almeida (UFSCar)

Prof(a) Dr(a) João Paulo Papa (UNESP)

Prof(a) Dr(a) André Backes (UFSCar)

O relatório de defesa assinado pelos membros da comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação