Journal of Intelligent & Robotic Systems (2025) 111:37
https://doi.org/10.1007/s10846-025-02220-9

REGULAR PAPER

Temporal Consistency as Pretext Task in Unsupervised Domain Adaptation for Semantic Segmentation

Felipe Barbosa[1] · **Fernando Osório**[1]

Received: 28 April 2024 / Accepted: 30 December 2024 / Published online: 19 March 2025

Abstract
Intelligent and autonomous robots (and vehicles) largely adopt computer vision systems to help in localization, navigation and obstacle avoidance tasks. By integrating different deep learning techniques, such as Object Detection and Image Semantic Segmentation, these systems achieve high accuracy in the domain they were trained on. Nonetheless, robustly operating in different domains still poses a major challenge to vision-based perception. In this sense, Unsupervised Domain Adaptation (UDA) has recently gained momentum due to its importance to real-world applications. Specifically, it leverages the prompt availability of unlabeled data to design auxiliary sources of supervision to guide the knowledge transfer between domains. The advantages of such an approach are two-fold: avoiding going through exhaustive labeling processes, and enhancing adaptation performance. In this scenario, exploring temporal correlations in unlabeled video data stands as an interesting alternative, which has not yet been explored to its full potential. In this work, we propose a Self-supervised learning framework that employs Temporal Consistency from unlabeled video sequences as a pretext task for improving UDA for Semantic Segmentation (UDASS). A simple yet effective strategy, it has shown promising results in a real-to-real adaptation setting. Our results and discussions are expected to benefit both new and experienced researchers on the subject.

Keywords Semantic segmentation · Unsupervised domain adaptation · Temporal consistency · Self-supervised learning · Review

1 Introduction

Intelligent and Autonomous Robots/Vehicles should be able to navigate in safe zones and avoid obstacles and dangerous zones. Therefore, it is very important for these systems to recognize the road (navigable zone), and the other elements present in the scene—"semantic elements" (e.g.: road, cars, pedestrians, trees, constructions, buildings, sidewalk, grass, animals, etc). Therefore, Semantic Segmentation is a task of utmost importance for visual perception in urban environments. It provides a summarized representation of a given scene, where elements are classified pixel-wise according to the set of categories under consideration.

✉ Felipe Barbosa
felipe.manfio.barbosa@usp.br

Fernando Osório
fosorio@icmc.usp.br

[1] Institute of Mathematics and Computer Science, University of São Paulo, São Paulo, Brazil

The field has historically evolved towards increasingly precise models, reaching Intersection over Union (IoU) values—the standard metric—of up to 90%. Nonetheless, these highly specialized models are prone to suffer with adapting to real-world scenarios, where the target data usually presents the so-called domain shift. This phenomenon is often caused by differences in appearance—illumination, textures, and so on—between the source domain the model was trained on and the target/application domain.

In this context, transfer-learning and fine-tuning techniques, usually associated with the presence of some sort of labels in the target domain, could be useful. However, the labeling process involves high human effort. This is even more critical for Semantic Segmentation tasks, which require dense labels—the "the curse of data labeling" [1]. Ultimately, it is impractical to obtain labeled data for all possible target domains.

In this sense, Unsupervised Domain Adaptation for Semantic Segmentation (UDASS) methods emerge as a promising new research direction, in the search for leveraging the promptly-available unlabeled data in domain adaptation.

Its practical relevance explains the increasing number of publications devoted to the subject.

Aligned with that, video streams are a great source of large amounts of unlabeled data. Despite that, temporal correlations among frames have rarely been explored in UDASS, thus leaving much room for improvements.

In light of that, we propose to explore Temporal Consistency in videos as a source of additional supervision to guide UDASS. On the one hand, it is simple to implement, since it does not require modifications to the base model's structure. On the other hand, precision and temporal stability can be simultaneously motivated in the target domain. Specifically, we aim at a cross-city real-to-real adaptation scenario, where such an approach has not yet been explored.

First, in Section 2, we conceptualize Domain Shift and (Unsupervised) Domain Adaptation. Section 3 compiles recent State-of-the-Art (SOTA) UDASS approaches that take into account temporal information from unlabeled video data. In Section 4, we present the proposed method. In Section 5 we share our findings from a real-to-real adaptation experiment, validating the employment of temporal data in UDASS. Finally, we draw our main conclusions in Section 6.

2 Domain Shift and Domain Adaptation

The field of Deep Learning has experienced large advances in the last decade, mainly fueled by the proposition of large annotated datasets [2–4]. Particularly, Semantic Segmentation is a well-developed research field, with recent contributions reaching up to 90% mean Intersection over Union (mIoU) in datasets such as Cityscapes [5].

However, the labeling process of such real-scenes datasets is labor-intensive: for example, the Cityscapes annotation took around 90 minutes per image. As an alternative to this scenario, a recent trend is to leverage synthetic data for model training. The main advantages of this approach are the possibility of simulating diverse scenarios, weather and illumination conditions, as well as sensor readings, all of that together with the associated labels.

Nonetheless, when trying to employ these models (trained on either real or synthetic data) in real-world applications, we will likely face a certain amount of performance degradation (Fig. 1). This can be caused by the so-called Domain Shift: differences between the source and target domains, such as illumination, textures, types of elements in the scenes, and so on. To deal with that, Domain Adaptation techniques try to transfer the knowledge from a given source domain to the target domain at hand.

To make the problem even worse, the adaptation process is not straightforward, since real-world target datasets usually lack annotations.

As a workaround, Unsupervised Domain Adaptation (UDA) was proposed to leverage the large availability of unlabeled data to boost the adaptation process without the need for labels.

According to the nature of source and target datasets, we can broadly define two categories of Domain Adaptation: synthetic-to-real, and real-to-real adaptation. In synthetic-to-real adaptation, synthetically-generated data are used during training, while the model is expected to run on real-world data. In real-to-real adaptation, both source and target datasets comprise real images, but with differences in appearance. Some authors group synthetic-to-real and real-to-real adaptation into Hard Domain Adaptation, while day-to-night, and season-to-season adaptation are often called Soft Domain Adaptation.

3 Related Works

The use of temporal information in UDASS is relatively new, with only a few contributions devoted to the theme. Nevertheless, interesting techniques and results have already been presented.

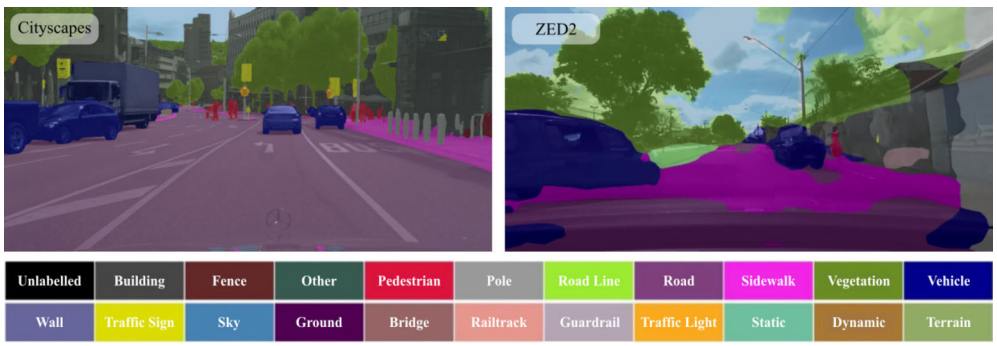

Fig. 1 Performance degradation caused by domain shift between source (Cityscapes) and target (ZED2) domains. Results considering a baseline Bisenet V2 model pretrained on the Cityscapes dataset

Fig. 2 Synthetic samples drawn from GTA5 (left) and SYNTHIA (right) datasets

The following sections present the systematic literature review developed to build the basis of our proposal and experimentation. Its main goal is to compile the latest advances in synthetic-to-real and real-to-real adaptation. We were particularly interested in mapping the use of video streams as a source of unlabeled data for UDASS. Therefore, we focused our search on this scope.

3.1 Synthetic-to-real Adaptation

The majority of works in UDASS address synthetic-to-real adaptation. In this scenario, the source domain is a synthetic dataset, usually SYNTHIA [6] or GTA5 [7] (Fig. 2), while the target dataset is composed of real urban scenes—Cityscapes being the common choice. Promising results (Table 1) have recently been shown by the adaptation techniques described in what follows.

Adversarial Learning is frequently employed in UDA setups. Broadly speaking, the main idea is to build a domain discriminator that is expected to become unable to distinguish between source and target domains. Ultimately, this means that the representations from both domains are aligned and, therefore, the domain shift has been overcome. DA-VSN [8] proposes a temporal consistency regularization (TCR) technique to minimize the divergence between source and target domains. Specifically, the authors design a Cross-domain Temporal Consistency Regularization (C-TCR) and an Intra-domain Temporal Consistency Regularization (I-TCR). Cross-domain motivates both spatial and spatiotemporal alignment between source and target predictions by means of two different discriminators (spatial and spatiotemporal). Intra-domain adaptation, conversely, is promoted through optical flow-based propagation of target frames, with subsequent conditioning of high-entropy predictions to have similar confidence to propagated low-entropy predictions. MoDA [9] learns target domain representations using self-supervised learning of object motion from unlabeled videos. Domain alignment of foreground and background classes is treated using different strategies. Foreground objects are tackled with the aid of instance-level guidance from the object motion. Background elements, on the other hand, are addressed with a category-specific discriminator in an adversarial strategy. Since motion learning works based solely on image correspondences, it is not affected by the cross-domain gaps.

Consistency Learning is another interesting approach to leveraging temporal information in UDASS. It can be explored in different configurations. TPS [10] proposes a consistency learning framework composed of cross-frame augmentation and cross-frame pseudo labeling. Through these mechanisms, the authors explore spatiotemporal cues for promoting prediction consistency between previous warped predictions, and current predictions computed from augmented frames. GUDA [11] explores photometric

Table 1 Current State-of-the-art in applying temporal data in the UDASS pipeline

Method Name	Year	Adaptation Strategy	Performance (mIoU %)	
			SYNTHIA → Cityscapes	GTA 5 → Cityscapes
CLST	2022	Contrastive Learning, Self-training	50.2 (16 classes) 58.1 (13 classes)	53.4
MoDA	2023	Adversarial, Multi-task Learning, Self-training, Consistency Learning	58.7 (16 classes) 68.3 (13 classes)	62.0
DA-VSN	2021	Adversarial, Consistency Learning	49.5 (11 classes)	
TPS	2022	Consistency Learning, Self-training	53.8 (11 classes)	
STPL	2023	Contrastive Learning, Self-training	51.8 (11 classes)	

Note: Missing values were not mentioned in the original papers

consistency in unlabeled videos by means of self-supervised learning of depth and ego-motion. These tasks can be formulated as a novel view synthesis problem, in which the target image is reconstructed using information from a reference image, given a predicted depth map and the relative transformation between the images. For that to be possible, the authors assume such transformation and camera intrinsic parameters known beforehand. In addition to that, semantic and surface normal loss are used to compose the proposed framework.

Contrastive Learning also shows promising results. In this line of research, class centroids are extracted so that to guide the domain alignment. The process is performed through contrastively learning to map features close to their corresponding centroids, and far away from the other classes' centroids. CLST [12] proposes to use contrastive learning to adapt category-wise centroids across domains. Source domain's class centroids are computed and online updated to act as the goal for alignment. This strategy has led to interesting results; nonetheless, since source priors may not be the optimal estimate for aligning the target domain, some authors have explored the use of domain-agnostic priors to guide both domains toward a shared representation [13]. In Spatio-Temporal Pixel-Level (STPL) [14] contrastive learning, spatio-temporal feature fusion, and pixel-level contrastive learning are applied in Source-free Domain Adaptation.

Compared to the previous works, the proposed approach holds some advantages. First, adversarial methods tend to be computationally costly and challenging/unstable to train [15]. Compared to Consistency Learning methods, our approach is simpler, since it does not motivate consistency between augmented versions of an image, just between the frames themselves. It also does not assume known intrinsic camera properties nor involves the composition of a large number of tasks, which may be difficult to optimize. Finally, the proposed method does not depend on the centroids used in Contrastive Learning, which may not be representative enough for both domains. This ultimately limits the adaptation process and requires the acquisition of positive and negative samples every time, incurring higher computational costs.

3.2 Real-to-real Adaptation

Real-to-real adaptation is an equally important task, where both source and target domains comprise real images. However, as it is more difficult to manually obtain labels for real images, this paradigm is less frequently tackled in the literature. Despite all the following works exploring UDASS, to the best of our knowledge, there are still no contributions leveraging temporal data in real-to-real adaptation.

We consider that in the case of real-to-real adaptation, a classification based on the target subtask is better suited than that based on the specific techniques employed.

Cross-city adaptation is the most frequently addressed adaptation subtask. One of the first works on the subject, FCNs in the Wild [16] tackles cross-city adaptation by trying to adapt from the training to the validation subset of Cityscapes. In the work of [17], cross-city adaptation is explored with Cityscapes-to-Oxford Robotcar [18] dataset. [19] studies the adaptation from Cityscapes to NTHU [20]. AdaptSegNet [21] and MaxSquareLoss [22, 23], besides synthetic-to-real adaptation, explore real-to-real adaptation using the Cityscapes and Cross-City [24] datasets.

Another subcategory of real-to-real adaptation comprises clear-to-adverse weather adaptation. In Advent [25], clear-to-foggy adaptation is explored using the Cityscapes dataset. CDAC [26] explores domain adaptation at both attention and output levels, based on a transformer architecture. Source and target domains are Cityscapes and ACDC [27], respectively. In the work of [28], the recently proposed Segment Anything Model (SAM) [29] is employed for refining pseudo-labels in the self-training of UDASS methods. Their main contribution resides in the different pseudo-labels fusing strategies so that to compensate for erroneous predictions for small-area and rare classes. Real-to-real adaptation is explored from Cityscapes to ACDC datasets.

Finally, there are also works devoted to day-to-nighttime adaptation. SePiCo [30] explores real-to-real adaptation from Cityscapes to various datasets, including: Dark-Zurich [31], Nighttime Driving [32], BDD100k-night [33], and Foggy Cityscapes [34]. The authors of MIC [35] propose an input masking mechanism to enable a teacher-student framework to learn to reconstruct masked pseudo-labels from the remaining unmasked regions. Consistency learning and Self-training are employed. Both Cityscapes-to-Dark Zurick and Cityscapes-to-ACDC are explored for day-to-nighttime and clear-to-adverse weather adaptation, respectively.

We explore cross-city adaptation. However, unlike all previous work, we explore temporal information from the abundant unlabeled video frames to guide the adaptation process.

4 Proposed Method

The proposed method consists of a Self-supervised mechanism for learning Temporal Consistency from unlabeled videos. Besides having shown promising results (Table 1), this technique is straightforward and can be seamlessly integrated into our base architecture. For that to be possible, we adapt the widely-adopted IoU metric for building our self-supervised auxiliary objective.

In the context of autonomous navigation in urban scenes, the compromise between inference speed and accuracy is more crucial than the accuracy taken in isolation. Therefore, we choose a lightweight model as our base architecture (Fig. 3), described in the following.

Quantitative evaluation is performed based on Temporal Consistency and Shannon's Entropy. In addition to that, visual inspection is performed on the learned features (activation maps), uncertainty maps, and segmentation output to certify the segmentation quality.

4.1 Architecture

Real-time applications, such as autonomous vehicles, need methods that deliver fast inference at low cost. With that in mind, and going against the majority of works that build upon increasingly heavy models, we choose a lightweight architecture as our base model.

Specifically, we build our proposal on top of the BisenetV2 model [36]. It has a dual-branch architecture composed of semantics and detail branches, each with its own goal. The semantics branch aims at the extraction of meaningful features, with low resolution and high semantic value; it has higher depth and lower spatial dimensions. The detail branch, on the other hand, preserves higher feature dimensionality, but operates at lower depths; with that, it tries to preserve spatial information. Feature fusion is performed by an attention-like mechanism, just before going through the final classification. The BisenetV2 model preserves yet another difference from its predecessor Bisenet [37]: multiple auxiliary segmentation heads to improve feature extraction. Figure 3 illustrates the BisenetV2 architecture.

Our architecture was built upon MMSegmentation's [38] implementation of BiseNet V2. MMSegmentation is an open-source library specialized in Semantic Segmentation, that provides implementation and pretrained weights for several model architectures and datasets. Specifically, we adopted as our backbone the BiseNet V2 model with weights pretrained for $160k$ iterations on 1024×1024 Cityscapes images, using batches of 16 images.

Our main goal is to perform self-supervised learning of temporal consistency to improve domain adaptation.

Self-supervised learning involves exploring the data at hand to derive new representations to be used as auxiliary sources of supervision. Image inpainting, image reconstruction, jigsaw puzzle solving, sequence order prediction and verification are some examples. Another interesting example is self-training, which can also be thought of as a type of self-supervised learning, since the model iteratively learns the goal task by means of its own predictions, termed pseudo-labels.

Temporal Consistency builds upon self-training and measures how coherent consecutive predictions (pseudo-labels) are. That is, given consecutive frames at times $t-1$ and t ($\boldsymbol{x_{t-1}}$ and $\boldsymbol{x_t}$, respectively), we measure how similar are the predictions for frame $\boldsymbol{x_t}$ and for frame $\boldsymbol{x_{t-1}}$ propagated to time t ($\boldsymbol{x_{t-1\to t}}$). The propagation, also called warping, is performed by using the optical flow from frame $t-1$ to frame t, $\boldsymbol{o_{t-1\to t}}$— extracted with a FlowNet 2.0 [39] model.

For that, we add an auxiliary module for computing the self-supervised objective. Therefore, no changes to the base model are necessary. The only difference to the original architecture comprises the Total Loss used to perform model optimization, which is now composed by the combination of

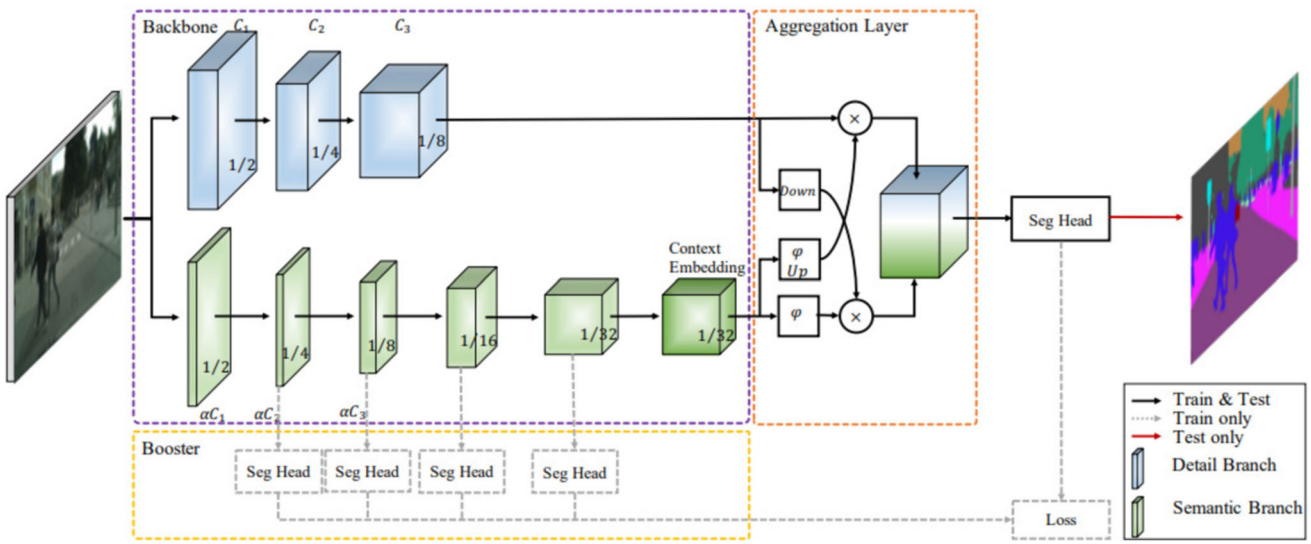

Fig. 3 Bisenet V2 [36] architecture

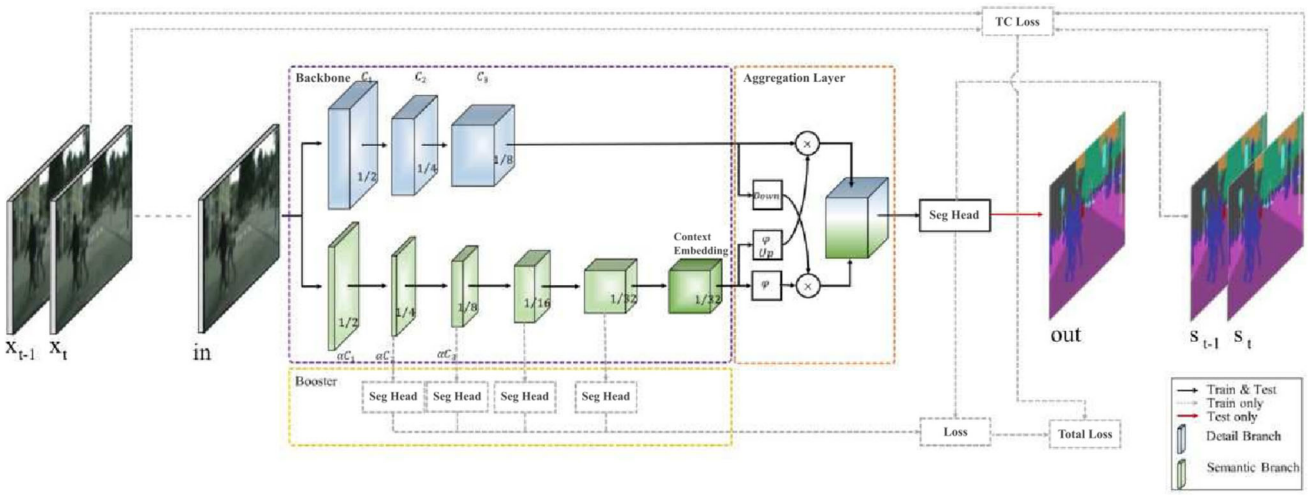

Fig. 4 Architecture proposed for performing UDASS by means of self-supervised temporal consistency learning

the Cross-Entropy Loss and the Temporal Consistency Loss. Our final architecture is illustrated in Fig. 4.

The advantages of such an approach are two-fold: first, no changes to the model itself are necessary; second, the auxiliary structures employed during training become non-operational during inference, so that the model preserves its inference speed.

4.2 Objective Function

Our main task is Semantic Segmentation, whose error is computed using the Cross-Entropy Loss (*SegLoss* in Eq. 1). It uses frames from the annotated source dataset and their corresponding predictions (*in* and *out* in Fig. 4), following the common training pipeline in supervised learning.

$$SegLoss = -\frac{1}{|I|}\sum_{i=1}^{I}\sum_{s=1}^{S} y_{n,s}\log(p_{n,s}) \tag{1}$$

The previous formulation is computed for an image I and a set of classes S, where $|I| = N \times X$ is the image dimensionality (number of pixels).

Our auxiliary objective, the Temporal Consistency Loss, receives consecutive frames ($\boldsymbol{x_{t-1}}$ and $\boldsymbol{x_t}$ in Fig. 4) drawn from the target domain, and their corresponding predictions ($\boldsymbol{y_{t-1}}$ and $\boldsymbol{y_t}$ in Fig. 4). It is implemented as an adaptation of the mIoU metric [40] ("Similarity Function" in Fig. 5) so that it operates on prediction probabilities, thus becoming differentiable (Eq. 2).

$$\widetilde{mIoU}(\boldsymbol{y'_t}, \tilde{\boldsymbol{y}}_t) = \frac{1}{|S|}\sum_{s\in S}\frac{\sum_{i\in I}|\boldsymbol{y'_{t,s,i}}\cdot\tilde{\boldsymbol{y}}_{t,s,i}|}{\sum_{i\in I}|\boldsymbol{y'_{t,s,i}}+\tilde{\boldsymbol{y}}_{t,s,i}-(\boldsymbol{y'_{t,s,i}}\cdot\tilde{\boldsymbol{y}}_{t,s,i})|} \tag{2}$$

In the previous formulation, $\tilde{\boldsymbol{y}}_t$ and $\boldsymbol{y'_t}$ are class probabilities, with $\boldsymbol{y'_t} = \tilde{\boldsymbol{y}}_{\boldsymbol{t-1\to t}}$ being the propagated probabilities from instant $t-1$ to instant t. Additionally, $I = H \cdot W$ is the number of pixels, and S corresponds to the set of classes being considered. The final Temporal Consistency Loss is computed as

$$TCLoss = 1 - \widetilde{mIoU_t} \tag{3}$$

The full pipeline of auxiliary loss computation is organized in the diagram from Fig. 5, and in Algorithm 1.

In Algorithm 1, *len* returns the number of samples in a given batch, and *warp* represents the prediction warping process performed based on the optical flow between consecutive input images. The variables $TC_{loss_{curr}}$ and $TC_{loss_{accum}}$ store the current and accumulated temporal consistency loss values, respectively.

The final loss is then computed as the weighted combination of the two losses (Eq. 4), where λ is a weighting factor

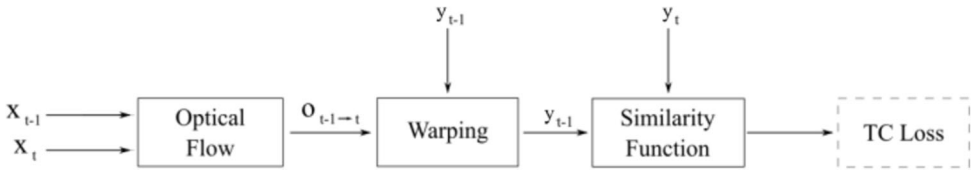

Fig. 5 Diagram illustrating the steps involved in calculating Temporal Consistency loss for auxiliary supervision

Algorithm 1 Calculate $TC_loss = 1 - \widetilde{mIoU}(\boldsymbol{y_{t-1}}, \boldsymbol{y_t})$.

Require: Segmentation model $f(\boldsymbol{x})$,
Optical flow model $of(\boldsymbol{x_{t-1}}, \boldsymbol{x_t})$,
Pair of sequential frames $\boldsymbol{x_{t-1}}$, and $\boldsymbol{x_t}$
Batch of samples

$num\_samples = len(batch)$
for x_{t-1}, x_t in batch **do**
  $\boldsymbol{o_{t-1 \to t}} \Leftarrow of(\boldsymbol{x_{t-1}}, \boldsymbol{x_t})$
  $\boldsymbol{y_{t-1}} \Leftarrow f(\boldsymbol{x_{t-1}})$
  $\boldsymbol{y_t} \Leftarrow f(\boldsymbol{x_t})$
  $\boldsymbol{y_{t-1 \to t}} \Leftarrow warp(\boldsymbol{y_{t-1}}, \boldsymbol{y_t}, \boldsymbol{o_{t-1 \to t}})$
  $TC_{loss_{curr}} \Leftarrow 1 - \widetilde{mIoU}(\boldsymbol{y_t}, \boldsymbol{y_{t-1 \to t}})$
  $TC_{loss_{accum}} \Leftarrow TC_{loss_{accum}} + TC_{loss_{curr}}$
end for

$TC_loss \Leftarrow TC_{loss_{accum}} / num\_samples$

for the Temporal Consistency loss.

$$Total\ Loss = SegLoss + \lambda \cdot TCLoss \tag{4}$$

5 Experiments and Results

This section presents our experimental setup in terms of the datasets used, the environment configuration, the training and inference parameters. Afterwards, in Section 5.4, we present the results obtained, followed by a fruitful discussion on challenges and open questions to UDASS, in Section 5.5.

The baseline performance for the method trained only on the source dataset is presented in Table 2.

5.1 Datasets

We adopt as source domain the widely-used Cityscapes dataset. Specifically, it comprises a set of German and Swiss urban scenes, from which we leverage the 5000 finely-annotated subset—obtained by labeling the 20th frame from each of its video snippets. The images were gathered at different times of the day, with clean and cloudy sky—good conditions. The dataset also provides a set of 20000 coarsely-annotated images, disparity maps, and associated camera parameters.

Our target domain, the ZED2 dataset [41], was captured by the authors with a ZED2 stereo camera. It comprises a set of videos (temporally-correlated frames and corresponding depth maps) from urban scenes in Brazilian streets, accounting for over 14, 000 frames. We performed data capture in good weather conditions: sunny and cloudy, at day time. No labels are provided. The choice for employing our own unlabeled data as the target domain can be explained by two aspects: first, it presents a domain shift from the source dataset, since it was captured by a different sensor and in

Table 2 Temporal Consistency results for all experimental configurations

Image Resolution	Weight	Baseline					Temporal Consistency					Improvement (%)				
		Road	Sidewark	Car	Pedestrian	Mean	Road	Sidewark	Car	Pedestrian	Mean	Road	Sidewark	Car	Pedestrian	Mean
256x256	0.3	0.837	0.710	0.487	0.088	0.333	0.911	0.785	0.477	0.529	0.523	8,84%	10,56%	-2,05%	501,14%	57,06%
	0.5						0.917	0.844	0.705	0.455	0.559	9,56%	18,87%	44,76%	417,05%	67,87%
	0.7						0.913	0.768	0.732	0.571	0.596	9,08%	8,17%	50,31%	548,86%	78,98%
512x512	0.3	0.828	0.633	0.557	0.162	0.363	0.924	0.825	0.826	0.280	0.612	11,59%	30,33%	48,29%	72,84%	68,60%
	0.5						0.912	0.878	0.874	0.886	0.741	10,14%	38,70%	56,91%	446,91%	104,13%
	0.7						0.884	0.875	0.648	0.906	0.808	6,76%	38,23%	16,34%	459,26%	122,59%
1024x1024	0.3	0.918	0.655	0.623	0.314	0.352	0.918	0.856	0.411	0.873	0.735	0,00%	30,69%	-34,03%	178,03%	108,81%
	0.5						0.913	0.857	0.758	0.895	0.844	-0,54%	30,84%	21,67%	185,03%	139,77%
	0.7						0.611	0.785	0.944	0.901	0.779	-33,44%	19,85%	51,52%	186,94%	121,31%

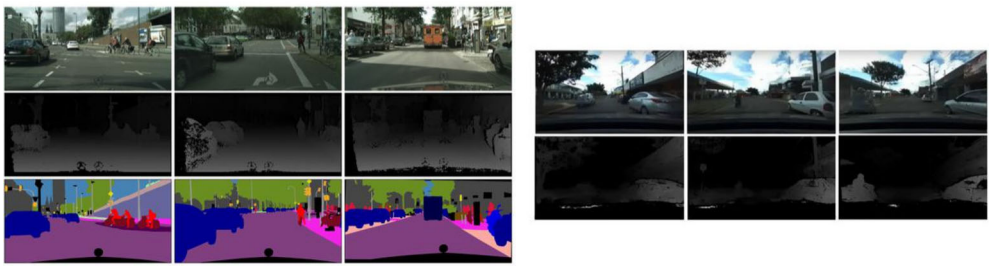

Fig. 6 Samples from the Cityscapes dataset (left), and from the dataset acquired with the ZED2 stereo camera (right)

an environment where elements and background differ in appearance with respect to the source domain; second, by employing this approach, we are closer to a real application scenario, where we should adapt a pretrained model to a new target domain where no labels are available.

Figure 6 illustrates examples drawn from both datasets.

5.2 Setup and Training Parameters

Following [42], we employed 30 epochs of fine-tuning starting from Cityscapes pretrained weights. Stochastic Gradient Descent (SGD) was employed as optimizer, with *momentum* of 0.9 and weight decay of 10^{-4}. The initial learning rate was set to 10^{-2}, and we adopted a polynomial update policy defined by $(1 - \frac{iter}{iters_{max}})^p$, with $p = 0.9$.

Random crop, random flip, photometric distortion, and padding were used as data-augmentation strategies. In all experiments, batch size was set to 4.

In total, 9 weight-resolution configurations were explored. That is, different values were tested for the weighting factor λ of the Temporal Consistency loss term in Eq. 4, including 0.3, 0.5 and 0.7. Regarding the input resolutions, we experimented with images with 256×256, 512×512 and 1024×1024 pixels.

All implementations were based on Python and Pytorch. Model training and inference were performed on a Google Colab environment equipped with an NVIDIA A100 GPU.

5.3 Inference

Inference was performed on a demo video from the ZED2 dataset, composed of 500 frames. Image resolution was set to 512×1024, and no data-augmentation strategies were employed. Results were qualitatively evaluated based on the Temporal Consistency and Entropy metrics. Visual inspection was performed on the activations from the last encoder layer, uncertainty maps, and the model outputs.

Visual inspection was performed on the activation maps to evaluate the features learned by the model. The visual inspection of the outputs, in turn, allows for evaluating model accuracy since we do not have labels in the target domain, what prevents us from quantitatively evaluating model precision—which would require a ground-truth.

The Temporal Consistency metric is defined as the mean Intersection Over Union between the prediction at time t ($\boldsymbol{y_t}$) and the prediction at time $t-1$ translated to time t ($\boldsymbol{y_{t-1\to t}}$) (5).

$$mIoU(\boldsymbol{y_{t-1\to t}}, \boldsymbol{y_t}) = \frac{1}{|S|} \sum_{s \in S} \frac{TP_s}{TP_s + FP_s + FN_s} \tag{5}$$

The prediction at time t is taken as the ground-truth, and TP_s, FP_s, and FN_s are the number of true positives, false positives and false negatives.

Since the Temporal Consistency term is operational only during training, no computational burden is added during inference, preserving the original model performance.

5.4 Results

The results from our approach are presented in the following sections.

Given that we explore Self-supervised learning on a completely unlabeled target dataset, no supervised metrics are possible to be calculated. Therefore, we rely on different strategies for inspecting and validating model performance, including hypothesis tests of Temporal Consistency and model uncertainty, as well as visual inspection of model predictions, uncertainty maps and the features learned by the network.

5.4.1 Temporal Consistency

Our experiments highlight that increased image resolution and loss weight lead to the highest gains in Temporal Consistency (Table 2).

Nonetheless, as it can be seen in Fig. 7, such gains are not necessarily related to gains in accuracy. The best results were actually observed for the lowest loss weight and image

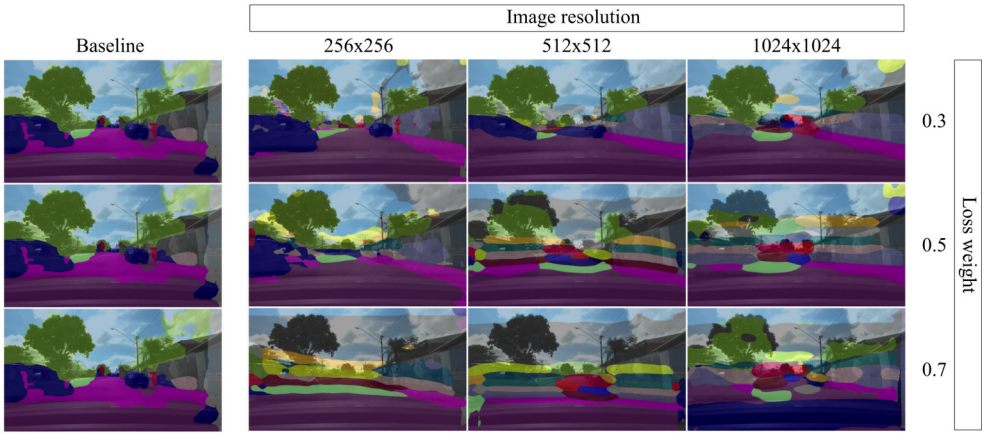

Fig. 7 Results from all experimental configurations. We clearly observe improvements derived from the use of unlabeled data from the target domain. However, higher loss weights and image resolutions lead to performance degradation

resolution: there is a clear improvement in the segmentation of navigable (street) and non-navigable (sidewalk) regions, which are critical for autonomous navigation in urban scenes. Classes such as wall, vegetation, buildings and vehicles also showed improvements.

This behavior can be understood by considering the nature of our Temporal Consistency Loss. Intuitively, static objects covering large regions are preferred, since they do not fluctuate to a greater extent over time. In fact, this is one of the major downsides of such an approach, since classes characterized by being small and thin—traffic signs (Fig. 8)—are prone to be disregarded as higher importance is given to the Temporal Consistency component of the loss. In this regard, leveraging model uncertainty, or class frequency information from the source domain, can be studied as alternatives to this problem.

Besides that, an interesting finding relates to the improvement of the overall temporal stability of the predictions. Figure 9 show that segmentation masks delivered by the baseline method are highly unstable, losing track of critical elements, such as cars. Our results—weight 0.3 and 256×256 image resolution—are more consistent over time.

The statistical validation of the relevance of our results was performed using the Student's T-test on the Temporal

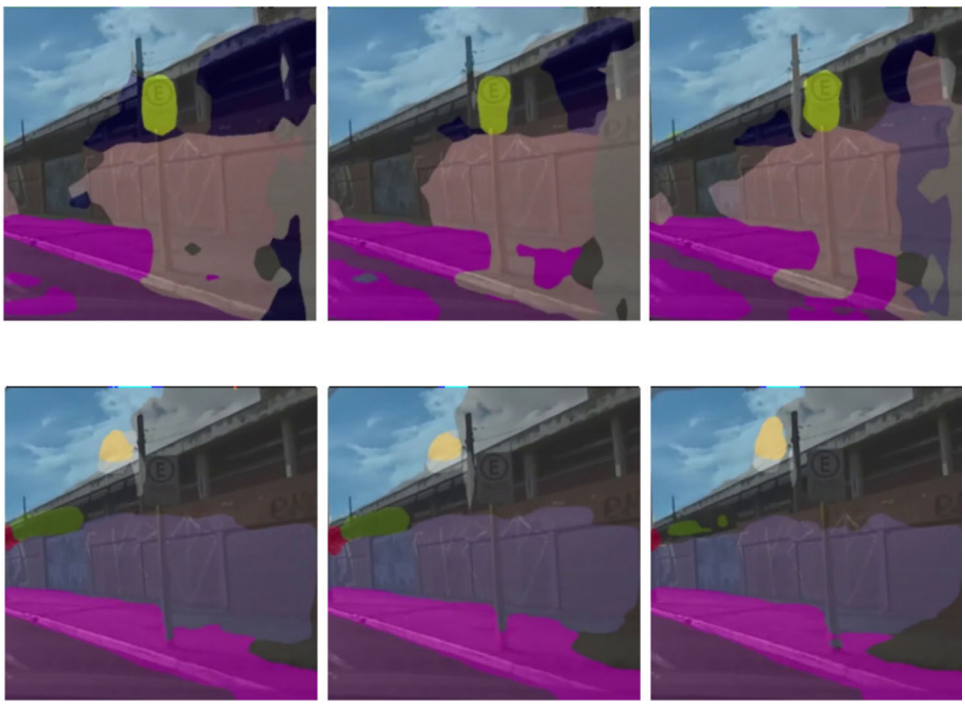

Fig. 8 Perception failures for small and thin objects in sequential frames. While the baseline model (upper row) identifies the traffic sign, ours (lower row) does not. Inherently, the Temporal Consistency loss encourages the presence of large static elements

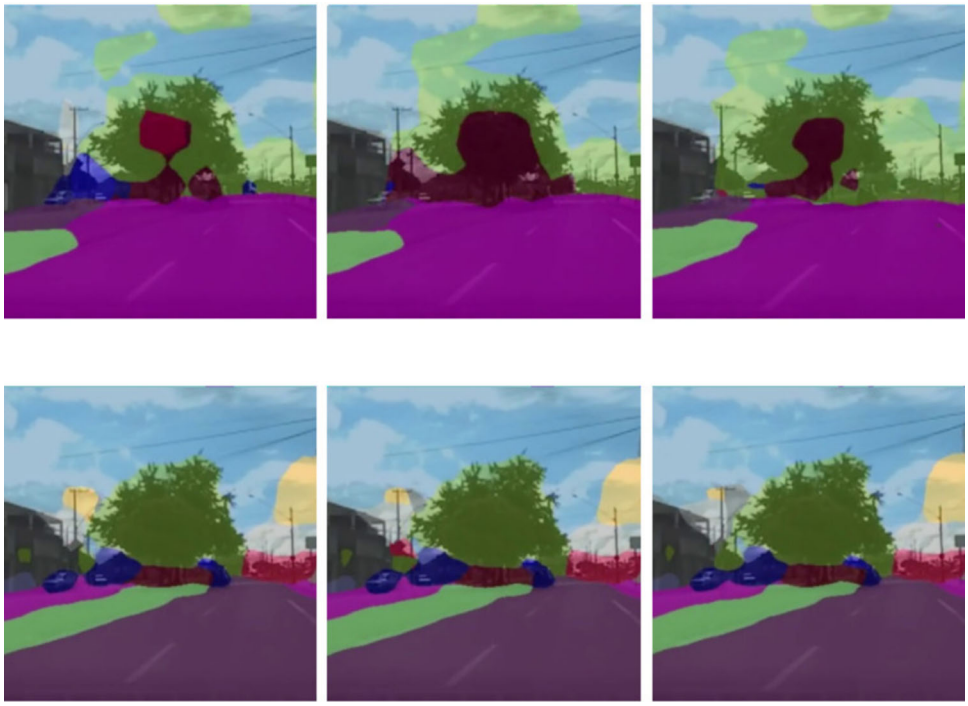

Fig. 9 Stability improvements in short-term windows. While the baseline model (first row) loses track of the vehicles in the image, our results (second row) are consistent over time

Consistency computed for the demo video snippet. The overall Temporal Consistency distribution of Baseline and Fine-tuned model is presented in Fig. 10.

We can observe that the Student's T'test assumptions are matched since both distributions follow a Gaussian Distribution. We can also observe that they are almost disjoint, meaning that the results found are relevant. This is confirmed by a t-statistic of -60.64, and a p-value of $4.26e-80$. That is, we can confidently reject the null hypothesis, which states that the means of the two distributions are equal.

We can also perform this analysis for class-wise results (Fig. 11). Considering critical classes for autonomous robots navigation, such as road, sidewalk, traffic light, and person, we observe a strong statistical relevance in the results obtained.

These findings illustrate that UDASS can be performed by leveraging Temporal Consistency in unlabeled video sequences drawn from the target domain. Thus, it stands as a promising alternative to computationally-heavy approaches, such as generative-adversarial setups.

However, it is also worth pointing out that the results are highly dependent on hyperparameter tuning, since majority of the configurations tested showed performance degradation in the target domain.

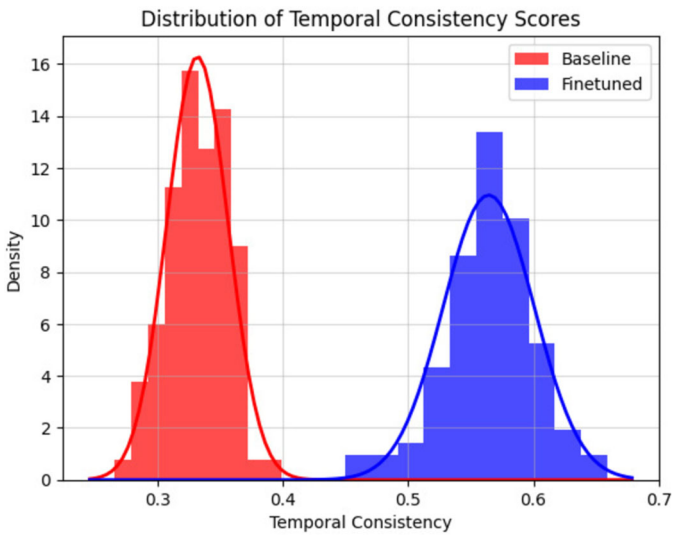

Fig. 10 Distribution of Temporal Consistency values for a demo video snippet from the target domain

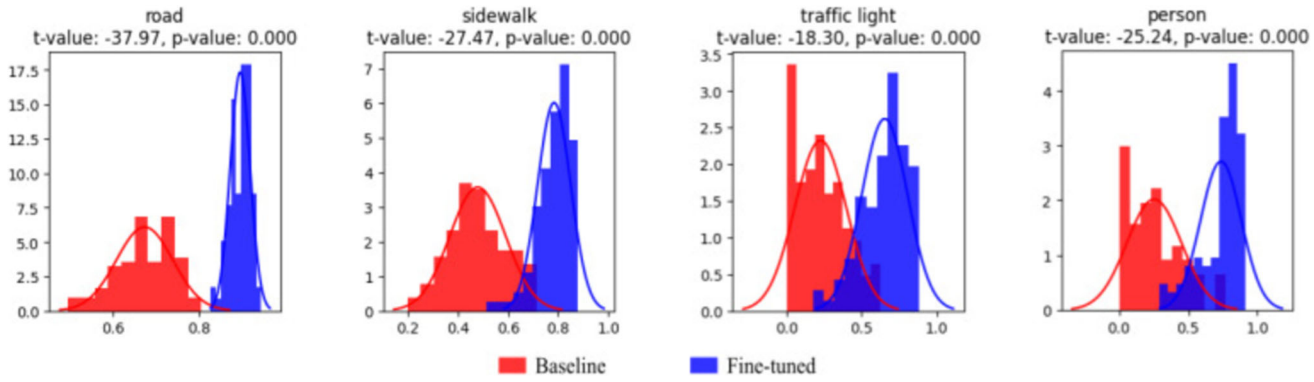

Fig. 11 Class-wise distribution of Temporal Consistency values for a demo video snippet from the target domain

5.4.2 Learned Features

Another commonly used analysis tool is the evaluation of intermediate activation maps from key layers. This allows us to verify where the model is focusing to make its predictions. Therefore, we can have a better understanding of which features in the input are the most representative for the model's current performance.

In our setup, the key layer was chosen as the output from the Aggregation Layer (Fig. 4).

Figure 4 illustrates activation maps extracted from the previously mentioned layer.

The upper row of Fig. 12 illustrates that the activations from baseline method were scattered over the image. From the lower row, we can observe that the Self-supervised learning of Temporal Consistency helped the activation maps to more consistently focus on critical regions for autonomous navigation, such as the road and the sidewalk.

5.4.3 Output Uncertainty

In self-supervised learning setups, output uncertainty is a useful measure of accuracy. In this sense, Shannon Entropy is widely adopted as a measure of model uncertainty and can be calculated by the formulation from Eq. 6.

$$H(X) = -\sum_{i=1}^{N} p_i \log p_i \tag{6}$$

The more uncertain the model predictions (ultimately all class probabilities equal), the higher the Shannon Entropy's value.

When comparing baseline and finetuned models, Fig. 13 shows a significant improvement of overall model uncertainty.

Adopting once again the Student's T'test, we obtain a t-statistic of 15.74, and a p-value of $1.06e-28$. That is, we

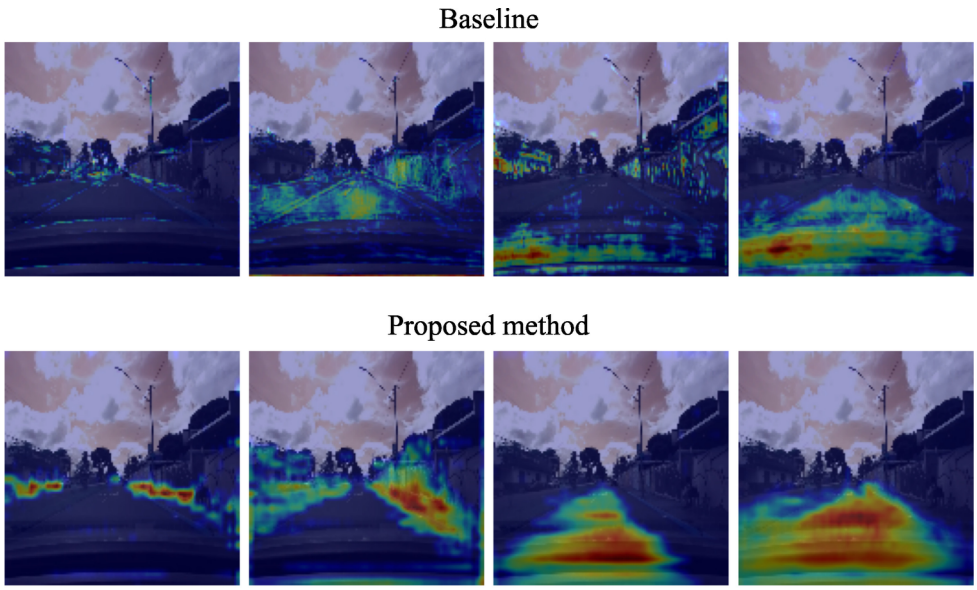

Fig. 12 Activation maps from the Aggregation Layer. Employing Self-supervised learning of Temporal Consistency helped improve the focus of the network to critical regions for autonomous navigation, such as road and sidewalk

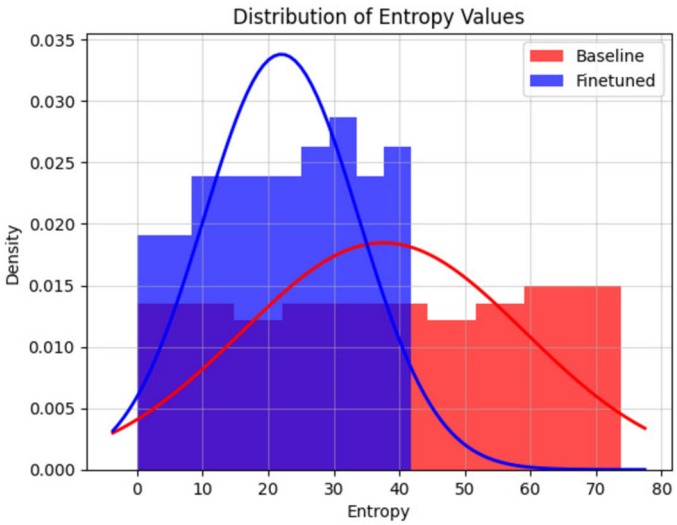

Fig. 13 Distribution of uncertainty values (Shannon Entropy) for a demo video snippet from the target domain

can confidently reject the null hypothesis, which states that the means of the two distributions are equal. Therefore, the reduction of model uncertainty is statistically relevant.

We can also perform this analysis for class-wise results (Fig. 14). Considering critical classes for autonomous robots navigation, such as road, sidewalk, traffic light, and person, we observe a strong statistical relevance in the results.

Uncertainty maps can also give us a visualization of the most challenging regions for the model (Fig. 15), i.e. the regions for which the predictions were uncertain.

As per Fig. 15, we can observe that the proposed approach mainly helped to reduce model uncertainty for classes with strong geometric priors, such as road, sidewalk and wall, which are also less prone to present strong fluctuations in their optical flow.

Most of the prediction uncertainty now lies in the class boundaries, which is expected given that perfectly defining class limits is challenging even for experienced human annotators.

The smoothing effect of the proposed method in the output uncertainty, however, leads to loss of precision for small and rare classes, such as pole and traffic sign. This behavior can be explained by the nature of the Temporal Consistency loss, which is more impacted by large static classes, such as road.

5.5 Discussion

Unsupervised Domain adaptation is a vibrant research subject, with several interesting results already demonstrated. However, there are still many challenges and open questions.

Despite frequent in the literature, global alignment delivers limited adaptation performance, not accounting for class specificities. In light of that, a recent trend in adversarial and contrastive domain adaptation is to perform class-wise adaptation.

In real-to-real alignment, it is still difficult to evaluate model performance in completely unlabeled target datasets. Feature alignment and output uncertainty can help to validate model pseudo-labels. Nevertheless, none of them is optimal: feature alignment is subject to interpretation, and low uncertainty does not necessarily mean high accuracy.

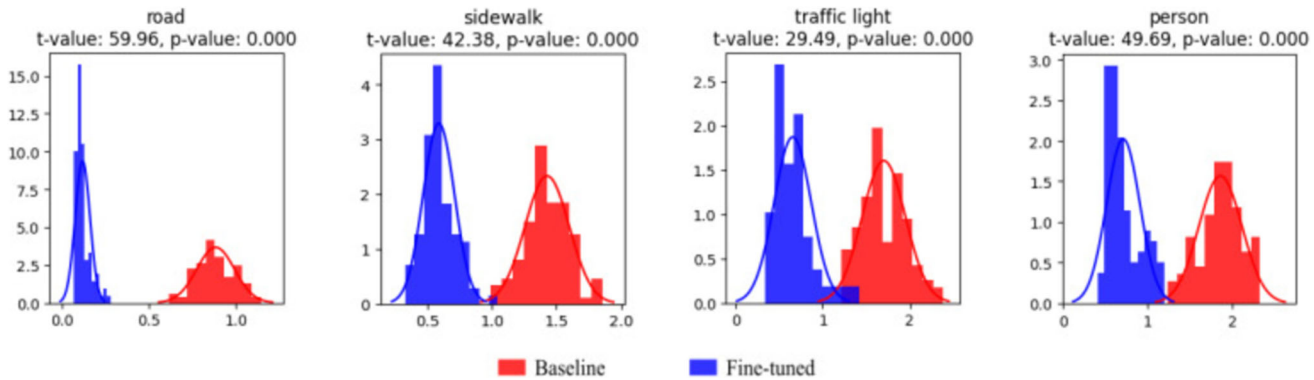

Fig. 14 Class-wise distribution of uncertainty values (Shannon Entropy) for a demo video snippet from the target domain

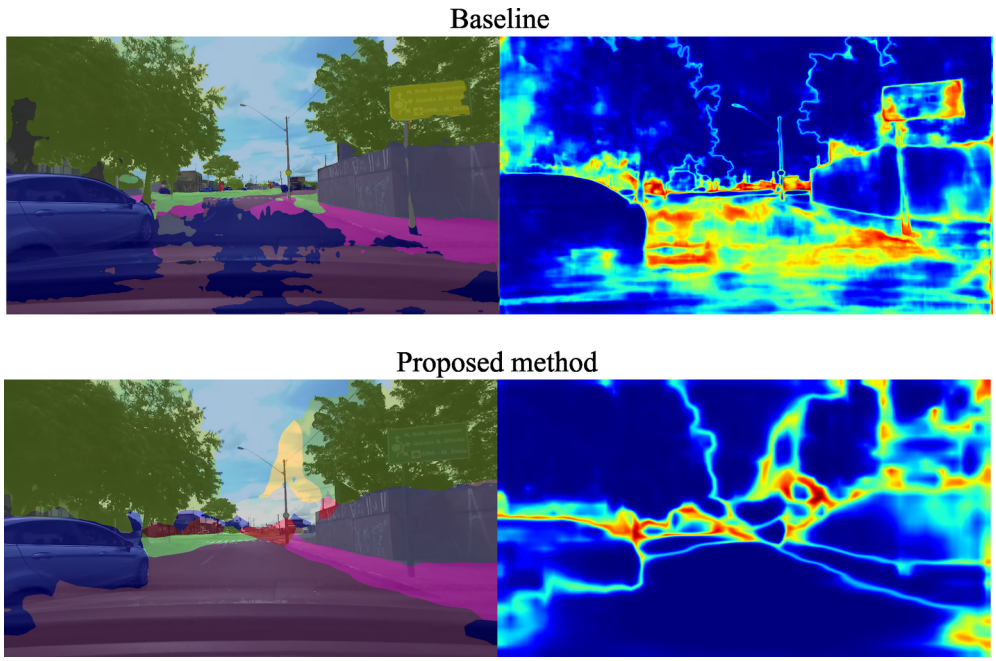

Fig. 15 Uncertainty maps highlight challenging regions for prediction. The proposed approach helped reduce model uncertainty for classes featuring strong geometric priors, such as road and sidewalk

Most of the current contributions are concerned with improving accuracy. However, real-world applications require faster inference under limited computational resources while preserving as much accuracy as possible.

The use of temporal data for UDASS also has several challenges. First, dealing with multiple frames at a time limits the batch size and input resolution. In addition, reproducibility is a major concern. Current contributions usually differ on the number of classes used for evaluation, not all hyperparameters are described and code is often not made publicly available.

Domain-agnostic adaptation has also recently been addressed. Researches on this subject try to generate a shared representation for both domains, which is not biased towards any of them. For instance, [13] generated class priors based on word embeddings describing each of the classes.

Exploring multiple data modalities can also be a very interesting path. Multi-modal models try to generate more robust shared representations to enhance adaptation performance. Depth data is an example that has been employed in multi-modal domain adaptation because of its strong geometric information, which suffers less from domain shift than the RGB images themselves.

Lastly, Open-set Domain Adaptation (source and target classes are not the same), Few-shot Domain Adaptation (adaptation performed with very few data), and Source-free domain adaptation (only the target dataset is available for training) are all promising directions with enormous practical relevance to real-world applications.

6 Conclusion

Domain Adaptation for Semantic Segmentation is a very exciting research field, with elevated practical relevance for autonomous mobile robots such as autonomous vehicles. However, the high costs involved in data annotation make it difficult to have labeled datasets for all possible target domains. In light of that, Unsupervised Domain Adaptation tries to leverage the promptly-available unlabeled data to learn useful representations that ultimately enhance adaptation performance.

A particularly powerful source of large amounts of unlabeled information is video streams. Specifically, the temporal correlation between frames can generate powerful auxiliary supervisory signals.

In light of that, we proposed a Self-supervised mechanism for learning Temporal Consistency between neighboring frames in unlabeled videos to help boost UDASS. As we have shown, this is a promising strategy since it does not incur changes to the base model's structure, and has delivered promising results in terms of both precision and temporal consistency.

Despite its appeal, such an approach still finds limited application. Some reasons that can explain this scenario are the overload imposed by dealing with multiple frames at once during training, which has as side-effects the limitation of batch sizes and image resolution.

Nonetheless, we consider that finding ways to better explore temporal data in UDASS setups is a promising

research direction, as shown by our findings in a real-to-real adaptation experiment.

The subject still faces several challenges, delivers limited accuracy, and leaves several open questions, to which we invite the reader to contribute, so that to foster and develop this research field that carries such an enormous practical relevance.

Author Contributions All authors contributed to the study conception and design. Material preparation was performed by Fernando Osório and Felipe Barbosa. Data collection and analysis were performed by Felipe Barbosa. The first draft of the manuscript was written by Felipe Barbosa and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability The data used in this research is publicly available. Source dataset is accessible through the Cityscapes dataset website (Cityscapes Dataset - Semantic Understanding of Urban Street Scenes. Available at: https://www.cityscapes-dataset.com/). Target dataset is made publicly available on a Google Drive repository, as described in this manuscript's text.

Code Availability Code is made publicly available on the project's page (Unsupervised Domain Adaptation through the exploration of Temporal Consistency. Available at: https://bit.ly/github-tc-semseg) on Github.

Declarations

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

1. Xie, J., Kiefel, M., Sun, M.-T., Geiger, A.: Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer (2016). https://doi.org/10.1109/CVPR.2016.401
2. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
3. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014, pp. 740–755. Springer, Cham (2014)
4. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5122–5130 (2017). https://doi.org/10.1109/CVPR.2017.544
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213–3223 (2016). https://doi.org/10.1109/CVPR.2016.350
6. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3234–3243 (2016). https://doi.org/10.1109/CVPR.2016.352
7. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) European Conference on Computer Vision (ECCV). LNCS, vol. 9906, pp. 102–118. Springer, (2016)
8. Guan, D., Huang, J., Xiao, A., Lu, S.: Domain Adaptive Video Segmentation via Temporal Consistency Regularization (2021). https://doi.org/10.1109/ICCV48922.2021.00795
9. Pan, F., Yin, X., Lee, S., Yoon, S., Kweon, I.-S.: Moda: Leveraging motion priors from videos for advancing unsupervised domain adaptation in semantic segmentation. ArXiv **abs/2309.11711** (2023)
10. Xing, Y., Guan, D., Huang, J., Lu, S.: Domain adaptive video segmentation via temporal pseudo supervision. ArXiv **abs/2207.02372** (2022)
11. Guizilini, V., Li, J., Ambruş, R., Gaidon, A.: Geometric unsupervised domain adaptation for semantic segmentation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8517–8527 (2021). https://doi.org/10.1109/ICCV48922.2021.00842
12. Marsden, R.A., Bartler, A., Döbler, M., Yang, B.: Contrastive learning and self-training for unsupervised domain adaptation in semantic segmentation. In: 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2022). https://doi.org/10.1109/IJCNN55064.2022.9892322
13. Huo, X., Xie, L., Hu, H., Zhou, W., Li, H., Tian, Q.: Domain-agnostic prior for transfer semantic segmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7065–7075 (2022). https://doi.org/10.1109/CVPR52688.2022.00694
14. Lo, S.-Y., Oza, P., Chennupati, S., Galindo, A., Patel, V.M.: Spatio-Temporal Pixel-Level Contrastive Learning-based Source-Free Domain Adaptation for Video Semantic Segmentation (2023). https://doi.org/10.1109/CVPR52729.2023.01015
15. Wang, Z., Yu, M., Wei, Y., Feris, R., Xiong, J., Hwu, W.-m., Huang, T.S., Shi, H.: Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12632–12641 (2020). https://doi.org/10.1109/CVPR42600.2020.01265
16. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. ArXiv **abs/1612.02649** (2016)
17. Zheng, Z., Yang, Y.: Rectifying Pseudo Label Learning via Uncertainty Estimation for Domain Adaptive Semantic Segmentation (2020). https://doi.org/10.48550/arXiv.2003.03773
18. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 Year, 1000km: The Oxford RobotCar Dataset. The International Journal of Robotics Research (IJRR) 36(1), 3–15 (2017) https://doi.org/10.1177/0278364916679498.

https://arxiv.org/abs/http://ijr.sagepub.com/content/early/2016/11/28/0278364916679498.full.pdf+html
19. Zou, Y., Yu, Z., Vijaya Kumar, B.V.K., Wang, J.: Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-training. Springer, Cham (2018)
20. Chen, Y., Chen, W., Chen, Y., Tsai, B., Wang, Y., Sun, M.: No more discrimination: Cross city adaptation of road scene segmenters. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2011–2020. IEEE Computer Society, Los Alamitos, CA, USA (2017). https://doi.org/10.1109/ICCV.2017.220 . https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.220
21. Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7472–7481 (2018). https://doi.org/10.1109/CVPR.2018.00780
22. Chen, M., Xue, H., Cai, D.: Domain adaptation for semantic segmentation with maximum squares loss. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2090–2099 (2019). https://doi.org/10.1109/ICCV.2019.00218
23. Wang, H., Shen, T., Zhang, W., Duan, L.-Y., Mei, T.: Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) Computer Vision - ECCV 2020, pp. 642–659. Springer, Cham (2020)
24. Chen, Y.-H., Chen, W.-Y., Chen, Y.-T., Tsai, B.-C., Wang, Y.-C.F., Sun, M.: No more discrimination: Cross city adaptation of road scene segmenters. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2011–2020 (2017). https://doi.org/10.1109/ICCV.2017.220
25. Vu, T.-H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2512–2521 (2019). https://doi.org/10.1109/CVPR.2019.00262
26. Wang, K., Kim, D., Feris, R., Betke, M.: CDAC: Cross-domain Attention Consistency in Transformer for Domain Adaptive Semantic Segmentation (2023). https://doi.org/10.1109/ICCV51070.2023.01058
27. Sakaridis, C., Dai, D., Van Gool, L.: Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10745–10755 (2021). https://doi.org/10.1109/ICCV48922.2021.01059
28. Yan, W., Qian, Y., Zhuang, H., Wang, C., Yang, M.: SAM4UDASS: When SAM Meets Unsupervised Domain Adaptive Semantic Segmentation in Intelligent Vehicles (2023). https://doi.org/10.1109/TIV.2023.3344754
29. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R.: Segment anything. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3992–4003 (2023). https://doi.org/10.1109/ICCV51070.2023.00371
30. Xie, B., Li, S., Li, M., Liu, C.H., Huang, G., Wang, G.: Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **45**(7), 9004–9021 (2023). https://doi.org/10.1109/TPAMI.2023.3237740
31. Sakaridis, C., Dai, D., Van Gool, L.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7373–7382 (2019). https://doi.org/10.1109/ICCV.2019.00747
32. Dai, D., Gool, L.V.: Dark model adaptation: Semantic image segmentation from daytime to nighttime. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 3819–3824 (2018). https://doi.org/10.1109/ITSC.2018.8569387
33. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2633–2642 (2020). https://doi.org/10.1109/CVPR42600.2020.00271
34. Sakaridis, C., Dai, D., Hecker, S., Van Gool, L.: Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018, pp. 707–724. Springer, Cham (2018)
35. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: MIC: Masked Image Consistency for Context-Enhanced Domain Adaptation (2023). https://doi.org/10.1109/CVPR52729.2023.01128
36. Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. Int. J. Comput. Vision **129**(11), 3051–3068 (2021). https://doi.org/10.1007/s11263-021-01515-2
37. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018, pp. 334–349. Springer, Cham (2018)
38. Contributors, M.: MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark (2020). https://github.com/openmmlab/mmsegmentation
39. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1647–1655 (2017). https://doi.org/10.1109/CVPR.2017.179
40. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vision **111**(1), 98–136 (2015). https://doi.org/10.1007/s11263-014-0733-5
41. Barbosa, F., Osório, F.: ZED2 Dataset (2023). https://bit.ly/zed2dataset
42. Varghese, S., Gujamagadi, S., Klingner, M., Kapoor, N., Bär, A., Schneider, J.D., Maag, K., Schlicht, P., Hüger, F., Fingscheidt, T.: An unsupervised temporal consistency (tc) loss to improve the performance of semantic segmentation networks. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 12–20 (2021). https://doi.org/10.1109/CVPRW53098.2021.00010

Felipe Barbosa received the B.S. degree in computer engineering from University of São Paulo, São Paulo, Brazil, in 2021 and is currently pursuing the M.S. degree in computer science and computational mathematics at University of São Paulo, São Paulo, Brazil. He is the author of articles on RGB-D visual perception for obstacle and safe zones segmentation. His research interests include deep learning, computer vision, RGB-D and video perception, and autonomous mobile robotics.

Fernando Osório received the M.S. and B.S.degrees in Computer Science from Federal University of Rio Grande do Sul (UFRGS), in 1988 and 1991, respectively. He received the Ph.D. degree at the Institut National Polytechnique de Grenoble, INPG, France, (1998). He has been a professor at University of São Paulo (ICMC/USP) since 2008. His current research interests include Machine Learning, Autonomous Vehicles, Robotics, Intelligent Driver Assistance Systems, Computer Vision, Pattern Recognition, and Industrial Robots.