Monitoring Traffic Flows via Unsupervised Domain Adaptation

Luca Ciampi*, Claudio Gennaro* and Giuseppe Amato*
*Institute of Information Science and Technologies - National Research Council - Pisa, Italy

***Abstract*—Monitoring traffic flows in cities is crucial to improve urban mobility, and images are the best sensing modality to perceive and assess the flow of vehicles in large areas. However, current machine learning-based technologies using images hinge on large quantities of annotated data, preventing their scalability to city-scale as new cameras are added to the system. We propose a new methodology to design image-based vehicle density estimators with few labeled data via an unsupervised domain adaptation technique.**

I. Introduction

Traffic problems are always increasing, and tomorrow's cities can only indeed be smart if they enable Smart Mobility. This concept is becoming more critical since traffic congestion caused by the increasing number of people using different road infrastructures to travel anywhere is imposing extra costs that make all activities more expensive and put a damper on the development.

Smart Mobility applications such as smart parking and road traffic management are nowadays widely employed worldwide, making our cities more livable and bringing benefits to the cities, a better quality of our life, reducing costs, and improving the energy usage.

Images are probably the best sensing modality to perceive and assess the flow of vehicles in large areas. Like no other sensing mechanism, networks of city cameras can observe such large dimensions and simultaneously provide visual data to AI systems to extract relevant information from this deluge of data.

In this work, we propose a CNN-based system that can estimate traffic density and count the vehicles present in urban scenes directly on-board smart city cameras, analyzing the images captured by themselves.

Current systems address the counting problem as a supervised learning process. They fall in two main classes of methods: a) detection-based approaches [1]–[3] that try to identify and localize single instances of objects in the image and b)density-based techniques that rely on regression techniques to estimate a density map from the image, and where the final count is given by summing all pixel values [4]. Figure 1 illustrates the mapping of such regression. Concerning vehicle counting in urban spaces, where images are of low resolution, and most objects are partially occluded, density-based methods have a clear advantage on detection methods [5].

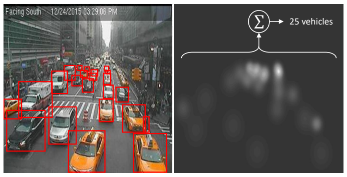

Fig. 1. Example of an image with the bounding box annotations (left) and the corresponding density map that sums to the counting value (right).

However, since this class of approaches requires pixel-level ground truth for supervised learning, they may not generalize well to unseen images, especially when there is a large *domain gap* between the training (*source*) and the test (*target*) sets, such as different camera perspectives, weather, or illumination. This gap severely hampers the application of counting methods to very large scale scenarios since annotating images for all the possible cases is unfeasible.

To mitigate this problem, we introduce a methodology that performs *unsupervised domain adaptation* among different scenarios, and we make publicly available two new datasets to conduct experiments. We evaluate our approach considering three different contexts: i) *Day2Night* domain adaptation, where the source domain is represented by images taken during the day and the target domain by pictures taken at night. ii) *Geometric* domain adaptation, where the source images belong to specific cameras, and the target ones are instead taken from different perspectives and contexts. iii) *Synthetic2Real* domain adaptation, where source images are collected using a video game and automatically annotated, while the target ones are real urban pictures. Experiments show a significant improvement compared to the performance of the model without domain adaptation.

II. The Datasets

In this section, we describe the datasets exploited in this work. In particular, an additional contribution of this work is creating two new datasets that we hope may be useful for other researchers in the future.

1) NDISPark Dataset: The *NDISPark - Night and Day Instance Segmented Park* dataset is a small, manually annotated dataset for counting cars in parking lots, consisting of about 250 images. This dataset is challenging and describes the most difficult situations that can be found in a real scenario: seven different cameras capture the images under various weather conditions and angles of view. Furthermore, it is worth noting

This work was partially supported by H2020 project AI4EU under GA 825619.

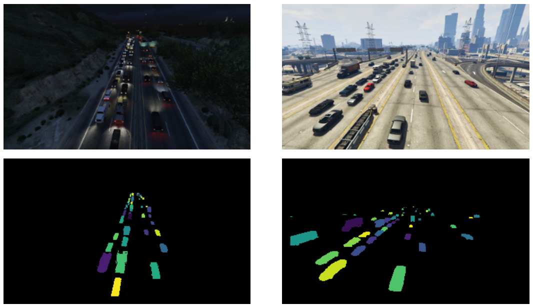

Fig. 2. Two examples of images of our *Grand Traffic Auto* dataset, together with the *automatically* generated instance segmentation annotations.

that pictures are taken during the day and the night, showing utterly different light conditions. The images are precisely annotated with *instance* segmentation labels, and this allowed us to generate accurate ground truth density maps usable for the counting task.

2) Grand Traffic Auto Dataset: The *GTA - Grand Traffic Auto* dataset is a vast collection of about 15,000 *synthetic* images of urban traffic scenes collected from the highly photo-realistic video game *GTA V - Grand Theft Auto V*. We deploy a framework that can *automatically* and precisely annotate the vehicles present in the scene with per-pixel annotations. To the best of our knowledge, it is the first *instance* segmentation synthetic dataset of city traffic scenarios. Figure 2 shows two examples of images belonging to this dataset together with the annotations.

3) WebCamT Dataset: The *WebCamT* dataset is a collection of traffic scenes recorded using city-cameras introduced by [6]. It is particularly challenging for analysis due to the low-resolution (352×240), high occlusion, and large perspective. We consider images belonging to different cameras and consequently having different views.

III. Proposed Method

Inspired by [7], we base our method on adversarial learning in the output space (i.e., the density maps), which contains valuable information such as scene layout and context. In our approach, we rely on the adversarial learning scheme to make the predicted density distributions of the source and target domains consistent.

The proposed framework consists of two modules: 1) a CNN that predicts traffic density maps and estimates the number of vehicles occurring in the scene, and 2) a discriminator that distinguishes whether the density map (received by the density map estimator) is generated processing an image of the source domain or the target domain.

In the training phase, the density map predictor learns to map images to densities, based on annotated data from the source domain. At the same time, it learns to fool the discriminator exploiting an adversarial loss, computed using the predicted density map of unlabeled images from the target domain. Consequently, we force the output space to have similar distributions for both the source and target domains.

IV. Experiments

We validate our approach using the three datasets described in the previous section and consequently considering three different scenarios:

- *Day2Night* Domain Adaptation: we employ the Night and Day Instance Segmented Park dataset, considering as source domain the images collected during the day, while as target domain the ones captured during the night.
- *Synthetic2Real* Domain Adaptation: we use the Grand Traffic Auto dataset, considering as source domain the synthetic images, while as target domain real traffic pictures.
- *Geometric* Domain Adaptation: we make use of the WebCamT dataset, considering as source domain images belonging to a set of cameras, and as target domain pictures belonging to a different set of perspectives and contexts.

Table IV shows the results of our approach compared against the model without the discriminator in terms of Mean Absolute Error (the lower is better).

	Day2Night	*Synth2Real*	*Geometric*
Baseline	3.95	4.10	3.24
Our Method	**3.49**	**3.88**	**2.86**

V. Conclusions

In this article, we tackle the problem of estimating the density and the number of vehicles present in large sets of urban traffic scenes. Building on a CNN-based density estimator, the proposed methodology can generalize to new sources of data for which there is no training data available. We achieve this generalization by adversarial learning, whereby a discriminator attached to the output induces similar density distribution in the target and source domains. Experiments show a significant improvement compared to the performance of the model without domain adaptation.

References

[1] G. Amato, L. Ciampi, F. Falchi, and C. Gennaro, "Counting vehicles with deep learning in onboard uav imagery," in *2019 IEEE Symposium on Computers and Communications (ISCC)*, 2019, pp. 1–6.

[2] L. Ciampi, G. Amato, F. Falchi, C. Gennaro, and F. Rabitti, "Counting vehicles with cameras," in *Proceedings of the 26th Italian Symposium on Advanced Database Systems, Castellaneta Marina (Taranto), Italy, June 24-27, 2018*, ser. CEUR Workshop Proceedings, vol. 2161. CEUR-WS.org, 2018.

[3] G. Amato, P. Bolettieri, D. Moroni, F. Carrara, L. Ciampi, G. Pieri, C. Gennaro, G. R. Leone, and C. Vairo, "A wireless smart camera network for parking monitoring," in *2018 IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–6.

[4] V. Lempitsky et al., "Learning to count objects in images," in *Advances in neural information processing systems*, 2010, pp. 1324–1332.

[5] Y. Li et al., "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1091–1100.

[6] S. Zhang et al., "Understanding traffic density from large-scale web camera data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5898–5907.

[7] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7472–7481.