Domain Adaptation for Traffic Density Estimation

Luca Ciampi[1][a], Carlos Santiago[2][b], Joao Paulo Costeira[2][c], Claudio Gennaro[1][d] and Giuseppe Amato[1][e]

[1] *Institute of Information Science and Technologies - National Research Council - Pisa, Italy*

[2] *Instituto Superior Técnico (LARSyS/IST), Lisbon, Portugal*

luca.ciampi@isti.cnr.it, carlos.santiago@tecnico.ulisboa.pt

Keywords: Unsupervised Domain Adaptation, Domain Adaptation, Synthetic Datasets, Deep Learning, Deep Learning for Visual Understanding, Counting Vehicles, Traffic Density Estimation, Convolutional Neural Networks

Abstract: Convolutional Neural Networks have produced state-of-the-art results for a multitude of computer vision tasks under supervised learning. However, the crux of these methods is the need for a massive amount of labeled data to guarantee that they generalize well to diverse testing scenarios. In many real-world applications, there is indeed a large *domain shift* between the distributions of the train (*source*) and test (*target*) domains, leading to a significant drop in performance at inference time. *Unsupervised Domain Adaptation* (UDA) is a class of techniques that aims to mitigate this drawback without the need for labeled data in the target domain. This makes it particularly useful for the tasks in which acquiring new labeled data is very expensive, such as for semantic and instance segmentation. In this work, we propose an end-to-end CNN-based UDA algorithm for traffic density estimation and counting, based on adversarial learning in the *output space*. The density estimation is one of those tasks requiring per-pixel annotated labels and, therefore, needs a lot of human effort. We conduct experiments considering different types of domain shifts, and we make publicly available two new datasets for the vehicle counting task that were also used for our tests. One of them, the *Grand Traffic Auto* dataset, is a *synthetic* collection of images, obtained using the graphical engine of the Grand Theft Auto video game, *automatically* annotated with precise per-pixel labels. Experiments show a significant improvement using our UDA algorithm compared to the model's performance without domain adaptation. The code, the models and the datasets are freely available at https://ciampluca.github.io/unsupervised_counting.

1 INTRODUCTION

With the advent of Convolutional Neural Networks (CNNs) (Lecun et al., 1998), supervised learning has reached excellent results across many Computer Vision application areas, such as object detection (Redmon and Farhadi, 2018) and instance segmentation (He et al., 2017). However, most CNN-based methods require a large amount of labeled data and make a common assumption: the training and testing data are drawn from the same distribution. The direct transfer of the learned features between different domains does not work very well because the distributions are different. Thus, a model trained on one domain, named *source*, usually experiences a drastic drop in performance when applied on another domain, named *target*. This problem is commonly referred as *Domain Shift* (Torralba and Efros, 2011).

Domain Adaptation is a common technique to address this problem. It adapts a trained neural network by fine-tuning it with a new set of labeled data belonging to the new distribution. However, in many real cases, gathering a further collection of labeled data is expensive, especially for tasks that imply per-pixel annotations, like semantic or instance segmentation.

Unsupervised Domain Adaptation (UDA) addresses the domain shift problem differently. It does not use labeled data from the target domain and relies only on supervision in the source domain. Specifically, UDA takes a source labeled dataset and a target *unlabeled* one. The challenge here is to automatically infer some knowledge from the target data to reduce the gap between the two domains.

In this work, we consider the counting task, defined as estimating the number of object instances

[a] https://orcid.org/0000-0002-6985-0439
[b] https://orcid.org/0000-0002-4737-0020
[c] https://orcid.org/0000-0001-6769-2935
[d] https://orcid.org/0000-0002-3715-149X
[e] https://orcid.org/0000-0003-0171-4315

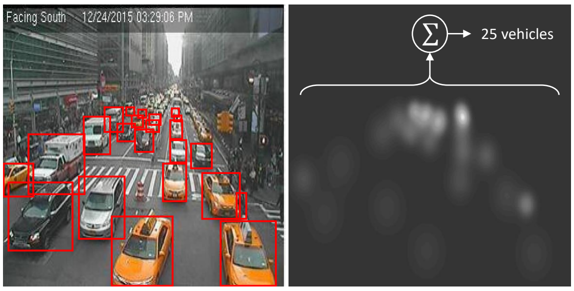


Figure 1: Example of an image with the bounding box annotations (left) and the corresponding density map that sums up to the counting value (right).

in still images or video frames (Lempitsky and Zisserman, 2010), which has recently attracted significant attention in the Computer Vision community. Specifically, we consider the vehicle counting scenario, where the task is to estimate the number of vehicles occurring in streets, roads, or parking lots. Most current systems address the counting task as a supervised learning process, relying on regression techniques to estimate a pixel-based density map from the image. The final count is obtained by summing all pixel values (Lempitsky and Zisserman, 2010). Figure 1 illustrates this approach.

We propose an end-to-end CNN-based UDA algorithm for traffic density estimation and counting, based on adversarial learning. Adversarial learning is performed directly on the generated density maps, i.e., in the *output space*, given that in this specific case, the output space contains valuable information such as scene layout and context. We focus on vehicle counting, but the approach is suitable for counting any other types of objects. To the best of our knowledge, we are the first to introduce a UDA scheme for counting to reduce the gap between the source and the target domain without using additional labels.

We conducted experiments considering different types of domain shifts and validating our approach on various vehicle counting datasets. First, we employed two existing datasets for traffic density estimation, *WebCamT* (Zhang et al., 2017a) and *TRANCOS* (Guerrero-Gómez-Olmedo et al., 2015). To emphasize the domain shift problem, we used as source domain images acquired by a specific subset of cameras. In contrast, we represented the target domain with images captured by a different subset of cameras, seeing different perspectives and visual contexts. We call this type of domain shift as the *Camera2Camera* domain shift. Comparisons with other techniques on these datasets show the superiority of our approach.

In order to test our technique with further types of domain shifts, we created and made publicly available the two additional datasets described in the following.

The *NDISPark - Night and Day Instance Segmented Park* dataset, consisting of images taken from surveillance cameras in a parking lot. Here, on the one hand, source data include annotated images collected by various cameras during the day. On the other hand, the unlabeled target domain contains images collected, in the same scenarios, during the night. We call this domain shift *Day2Night*.

The *GTA - Grand Traffic Auto* dataset, a vast collection of *synthetic* images generated with the highly photo-realistic graphical engine of the *Grand Theft Auto V* video game, developed by *Rockstar North*. This dataset consists of urban traffic scenes, *automatically* and precisely annotated with per-pixel annotations. To the best of our knowledge, it is the first *instance* segmentation synthetic dataset of traffic scenarios. We use this dataset to train the counting algorithm. Then, we performed domain adaptation to be able to count in real images. In this case, the domain shift is represented by the *Synthetic2Real* difference.

Figure 2 summarizes the described domain shifts that we have addressed.

In all the experiments, we show that our UDA technique always outperforms the non-domain adapted models.

Contributions of this work can be summarized as follows:

- We introduce a UDA algorithm for traffic density estimation and counting, which can reduce the domain gap between a labeled source dataset and an unlabeled target one. To the best of our knowledge, this is the first time that UDA is applied to counting.
- We create and make publicly available two new datasets, both having instance segmentation annotations. One is manually annotated and consists of images of parked cars collected during the day and by night. The second is a synthetic collection of images taken from a photo-realistic graphical engine, where the per-pixel annotations are automatically created.
- We conduct extensive experiments taking into account three different types of domain shifts and validating our technique on various vehicle counting datasets, demonstrating a significant improvement using our UDA algorithm compared to the model's performance without domain adaptation.

2 RELATED WORK

This section reviews some previous work related to the Unsupervised Domain Adaptation and the

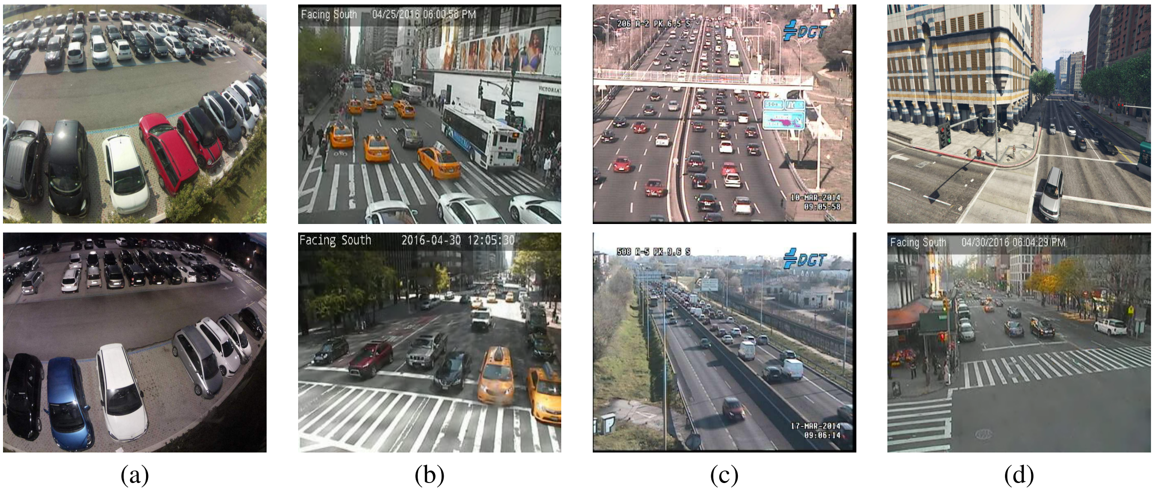

Facing South 04/25/2016 06:00:58 PM
Facing South 2016-04-30 12:05:30
DGT
DGT
Facing South 04/30/2016 06:04:29 PM
(a)
(b)
(c)
(d)

Figure 2: The Domain Shift scenarios that have been addressed in this work: (a) *Day2Night*; (b) and (c) *Camera2Camera*; (d) *Synthetic2Real*. The first row represents the labeled *source* domain, while the second represents the unlabeled *target* one used for our unsupervised domain adaptation.

Counting task.

2.1 Unsupervised Domain Adaptation

Traditional UDA approaches have been developed to address the problem of image classification, and they try to align features across the two domains ((Ganin and Lempitsky, 2015), (Tzeng et al., 2017)). However, as pointed out in (Zhang et al., 2017b), they do not perform well in other tasks.

More recent advances also involve the semantic segmentation task. In this case, adversarial training for UDA is the most employed approach. It includes two networks. The first predicts the segmentation maps for the input source image. The second acts as a discriminator, taking the feature maps from the segmentation network and trying to predict the input domain. The adversarial loss, computed from the discriminator output, tries to make the distributions of the two domains more similar. The first to apply such a technique is (Hoffman et al., 2016). More recently, the work proposed in (Hong et al., 2018) employs a residual network and adversarial training to make the source feature maps closer to the target ones. The authors of (Chen et al., 2019) combine semantic segmentation and depth estimation to boost the adaptation performance, providing to the discriminator the segmentation and the depth prediction maps jointly. Another interesting work that inspired this paper is (Tsai et al., 2018), where the authors applied adversarial training to the output space taking advantage of the structural consistency across domains.

A very appealing application of domain adaptation concerns synthetic data, which has led to the development of several synthetic datasets, such as ViPeD ((Amato et al., 2019), (Ciampi et al., 2020a)) for pedestrian detection and SYNTHIA (Ros et al., 2016) for semantic segmentation and autonomous driving applications. In this case, the algorithm is trained using these synthetic images and applied over real images. The domain adaptation algorithm is in charge of filling the gap between the two worlds.

2.2 The Counting Task

Following the taxonomy adopted in (Sindagi and Patel, 2018), we can broadly classify existing counting approaches into two categories: counting by detection and counting by regression. Counting by *detection* is a supervised technique where we localize instances of the objects, and then we count them. Some relevant works present in the literature are (Ciampi et al., 2018), (Amato et al., 2019), (Amato et al., 2018), (Aich and Stavness, 2018), (Laradji et al., 2018). Instead, Counting by *regression* (Lempitsky and Zisserman, 2010) is a supervised learning approach that tries to establish a direct mapping (linear or not) from the image features to the number of objects present in the scene or a corresponding density map (i.e., a continuous-valued function), skipping the challenging task of detecting instances of the objects.

Regression techniques have shown superior performance in crowded scenarios where the objects' instances are sometimes not clearly visible due to occlusions, and they have been applied to a multitude of situations. The first work that employed a pure CNN to estimate the density and count people in crowded contexts is presented by (Boominathan et al., 2016). A more efficient structure is proposed by (Zhang et al., 2016) introducing a Multi-Column CNN-based architecture (MCNN) for crowd counting. A similar idea is developed by (Oñoro-Rubio and López-Sastre, 2016) with a scale-aware, multi-column counting model named Hydra-CNN able to estimate traffic densities in congested scenes. More recently, the authors of (Li et al., 2018) introduced CSRNet. This CNN-based algorithm uses dilated kernels to deliver larger reception fields and replace pooling operations. We employ this network as the baseline in our work, and we briefly review its architecture in the next sections.

The main limitations of these approaches are due to the scarcity of data. As a result, existing methods often suffer from overfitting, which leads to performance degradation while transferring them to other scenes. Besides, there is another inherent problem: the labels of these datasets are not very accurate. Most of the existing datasets are dot-annotated. Consequently, the ground truth density maps are just an approximation in which the objects' sizes are estimated using some heuristics. This work addresses both problems proposing an unsupervised domain adaptation technique that exploits unlabeled data and introduces two new datasets with per-pixel annotations that allow the creation of precise ground truth density maps. To the best of our knowledge, this work is the first that employs UDA to the counting task, extending the very preliminary results obtained in (Ciampi et al., 2020b), where a similar approach was exploited in just one limited scenario.

3 DATASETS

As mentioned before, to prove our approach's validity, we performed experiments on various vehicle counting datasets, offering different domain shift characteristics. Specifically, we exploited two existing datasets for traffic density estimation: *WebCamT* (Zhang et al., 2017a) and *TRANCOS* (Guerrero-Gómez-Olmedo et al., 2015). Then, we used two additional datasets that we created on purpose and made publicly available: the *NDISPark - Night and Day Instance Segmented Park* dataset and the *GTA - Grand Traffic Auto* dataset. Figure 3 shows some images belonging to these datasets, together with the associated labels and the corresponding generated density maps used for the counting task. In the next sections, we describe more in detail each of them.

3.1 WebCamT Dataset

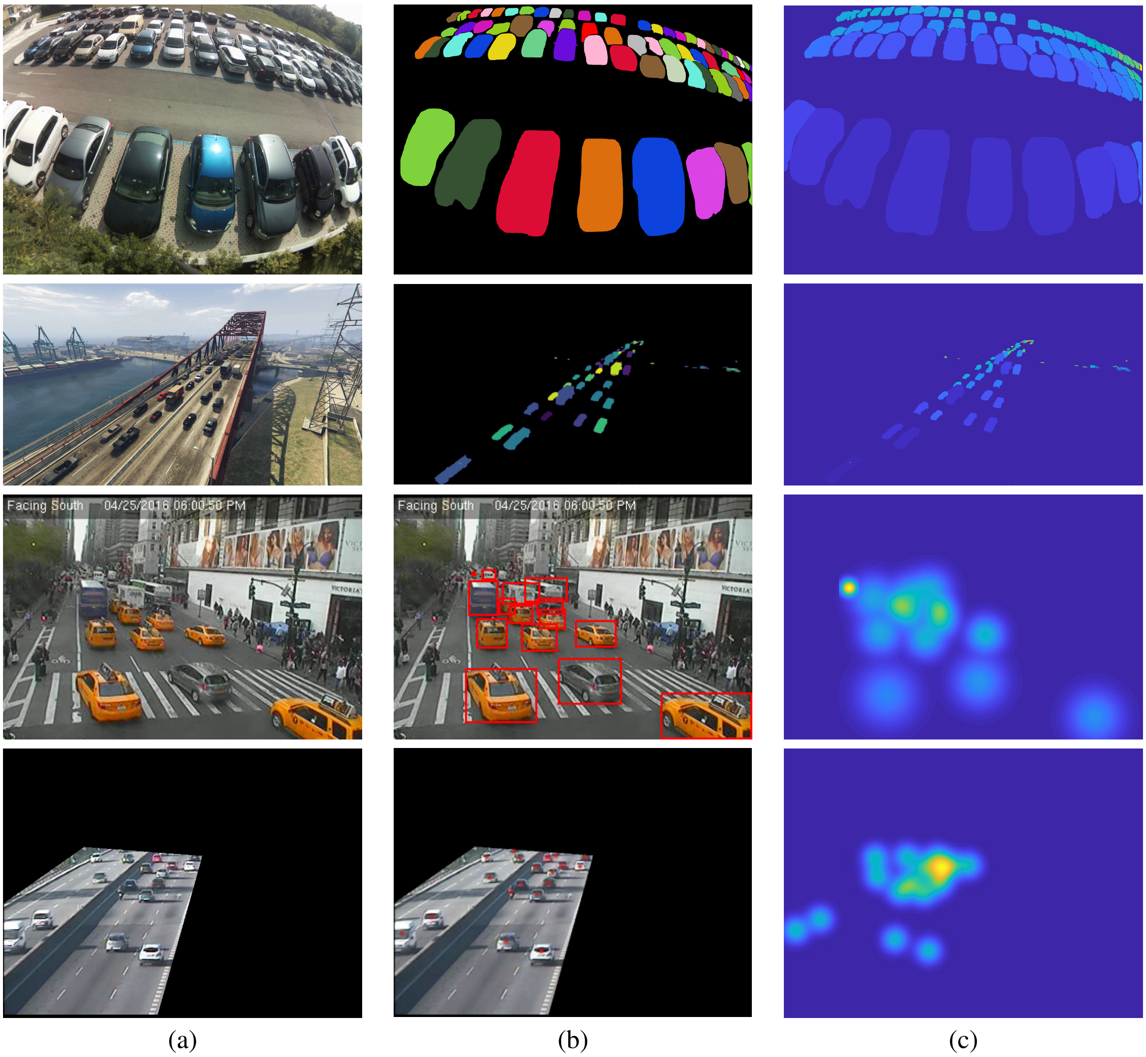
The *WebCamT* dataset is a collection of traffic scenes recorded using city-cameras introduced by (Zhang et al., 2017a). It is particularly challenging to analyze due to the low-resolution (352×240), high occlusion, and large perspective. We consider a total of about 40,000 images belonging to 10 different cameras and consequently having different views. We employ the existing bounding box annotations of the dataset to generate ground truth density maps. In particular, we consider one Gaussian Normal kernel for each vehicle present in the scene, having a value of μ and σ equal to the center and proportional to the size of the bounding box surrounding the vehicle, respectively. We used this dataset to test performance with the *Camera2Camera* domain shift, introduced in Section 1.

3.2 TRANCOS Dataset

The *TRANCOS* dataset is a public dataset containing 1244 dot-annotated images of different congested traffic scenes captured by surveillance cameras, introduced by (Guerrero-Gómez-Olmedo et al., 2015). The approximated ground truth density maps are generated by putting one Normal Gaussian kernel for each dot present in the scene, having a value of σ empirically decided by the authors. They also provided the regions of interest (ROIs) for each image. We used this dataset to test performance with the *Camera2Camera* domain shift, mentioned in Section 1.

3.3 NDISPark Dataset

The *NDISPark - Night and Day Instance Segmented Park* dataset was created by us on purpose and made publicly available. It is a small, manually annotated dataset for counting cars in parking lots, consisting of about 250 images. This dataset is challenging and describes most of the problematic situations that we can find in a real scenario: seven different cameras capture the images under various weather conditions and viewing angles. Another challenging aspect is the presence of partial occlusion patterns in many scenes such as obstacles (trees, lampposts, other cars) and shadowed cars. Furthermore, it is worth noting that images are taken during the day and the night, showing utterly different lighting conditions and that, unlike most counting datasets, the *NDISPark* dataset is precisely annotated with *instance* segmentation labels, allowing us to generate accurate ground truth density maps for the counting task since the size of the vehicles is well-known. We employed this dataset


Facing South 04/25/2016 06:00:50 PM
Facing South 04/25/2016 06:00:50 PM

(a) (b) (c)

Figure 3: Some examples taken by the four datasets used in this work: (a) Images; (b) Labels; (c) Density Maps generated from the labels. Each row correspond to a specific dataset: from top to bottom, the *NDISPark - Night and Day Instance Segmented Park* and the *GTA - Grand Traffic Auto* datasets introduced in this work, the *WebCamT* dataset (Zhang et al., 2017a) and the *TRANCOS* dataset (Guerrero-Gómez-Olmedo et al., 2015). Note that the densities maps generated in our datasets are accurate since we start from an instance segmentation annotations. Also notice that, in the case of the *GTA - Grand Traffic Auto* dataset, annotations are *automatically* generated without human effort.

to test performance with the *Day2Night* domain shift, explained in Section 1.

3.4 GTA Dataset

The *GTA - Grand Traffic Auto* dataset was also created by us on purpose and made publicly available. It is a vast collection of about 15,000 *synthetic* images of urban traffic scenes collected using the highly photo-realistic graphical engine of the *GTA V - Grand Theft Auto V* video game. About half of them concern urban city areas, while the remaining involve sub-urban areas and highways. To generate this dataset, we designed a framework that *automatically* and precisely annotates the vehicles present in the scene with per-pixel annotations. To the best of our knowledge, this is the first *instance* segmentation synthetic dataset of city traffic scenarios. As in the *NDISPark* dataset, the instance segmentation labels allow us to produce accurate ground truth density maps for the counting task since the size of the vehicles is well-known. We exploited this dataset to test performance with the *Synthetic2Real* domain shift, introduced in Section 1.

4 PROPOSED METHOD

Our method relies on a CNN model trained end-to-end with adversarial learning in the output space (i.e., the density maps), which contains rich information such as scene layout and context. The peculiarity of our adversarial learning scheme is that it forces the predicted density maps in the target domain to have local similarities with the ones in the source domain.

Figure 4 depicts the proposed framework consisting of two modules: 1) a CNN that predicts traffic density maps, from which we estimate the number of vehicles in the scene, and 2) a discriminator that identifies whether a density map (received by the density map estimator) was generated from 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 predict realistic density maps for the target domain by trying to fool the discriminator with an adversarial loss. The discriminator's output is a pixel-wise classification of a low-resolution map, as illustrated in Figure 4, where each pixel corresponds to a small region in the density map. Consequently, the output space is forced to be locally similar for both the source and target domains. In the inference phase, the discriminator is discarded, and only the density map predictor is used for the target images. We describe each module and how it is trained in the following subsections.

4.1 Density Estimation Network

We formulate the counting task as a density map estimation problem (Lempitsky and Zisserman, 2010). The density (intensity) of each pixel in the map depends on its proximity to a vehicle centroid and the size of the vehicle in the image so that each vehicle contributes with a total value of 1 to the map. Therefore, it provides statistical information about the vehicles' location and allows the counting to be estimated by summing of all density values.

This task is performed by a CNN-based model, whose goal is to automatically determine the vehicle density map associated with a given input image. Formally, the density map estimator, $\Psi : \mathcal{R}^{C\times\mathcal{H}\times\mathcal{W}} \mapsto \mathcal{R}^{\mathcal{H}\times\mathcal{W}}$, transforms a $\mathcal{W}\times\mathcal{H}$ input image I with $\mathcal{C}$ channels, into a density map, $D = \Psi(I) \in \mathcal{R}^{\mathcal{H}\times\mathcal{W}}$.

4.2 Discriminator Network

The discriminator network, denoted by Θ, also consists of a CNN model. It takes as input the density map, D, estimated by the network Ψ. Its output is a lower resolution probability map where each pixel represents the probability that the corresponding region (from the input density map) comes either from the source or the target domain. The goal of the discriminator is to learn to distinguish between density maps belonging to source or target domains. Through an adversarial loss, this discriminator will, in turn, force the density estimator to provide density maps with similar distributions in both domains. In other words, the target domain density maps have to look realistic, even though the network Ψ was not trained with an annotated training set from that domain.

4.3 Domain Adaptation Learning

The proposed framework is trained based on an alternate optimization of the density estimation network, Ψ, and the discriminator network, Θ. Regarding the former, the training process relies on two components: 1) density estimation using pairs of images and ground truth density maps, which we assume are only available in the source domain; and 2) adversarial training, which aims to make the discriminator fail to distinguish between the source and target domains. As for the latter, images from both domains are used to train the discriminator on correctly classifying each pixel of the probability map as either source or target.

To implement the above training procedure, we use two loss functions: one is employed in the first step of the algorithm to train network Ψ, and the other is used in the second step to train the discriminator Θ. These loss functions are detailed next.

Network Ψ Training. We formulate the loss function for Ψ as the sum of two main components:

$$\mathcal{L}(I^{\mathcal{S}}, I^{\mathcal{T}}) = \mathcal{L}_{density}(I^{\mathcal{S}}) + \lambda_{adv}\mathcal{L}_{adv}(I^{\mathcal{T}}), \qquad (1)$$

where $\mathcal{L}_{density}$ is the loss computed using ground truth annotations available in the source domain, while $\mathcal{L}_{adv}$ is the adversarial loss that is responsible for making the distribution of the target and the source domain closer to each other. In particular, we define the density loss $\mathcal{L}_{density}$ as the mean square error between the predicted and ground truth density maps, i.e. $\mathcal{L}_{density} = MSE(D^{\mathcal{S}}, D^{\mathcal{S}\text{-}\mathcal{GT}})$.

To compute the adversarial loss $\mathcal{L}_{adv}$, we first forward the images belonging to the target domain through network Ψ, to generate the predicted density maps $D^{\mathcal{T}}$. Then, we forward $D^{\mathcal{T}}$ through network Θ, to generate the probability map $P = \Theta(\Psi(I^{\mathcal{T}})) \in [0,1]^{H'\times W'}$, where $H' < H$ and $W' < W$. The adversarial loss is given by

$$\mathcal{L}_{adv}(I^{\mathcal{T}}) = -\sum_{h,w} \log(P_{h,w}), \qquad (2)$$

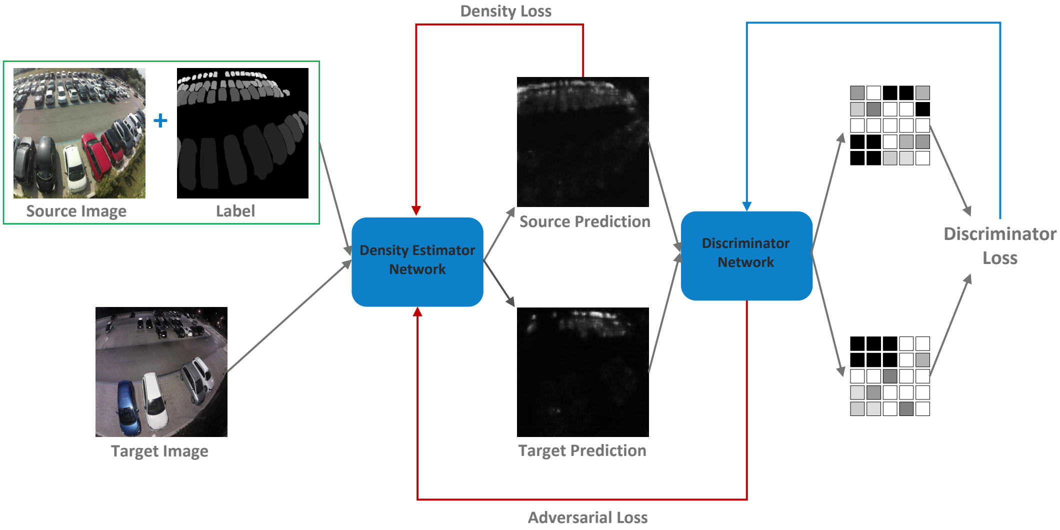


Figure 4: Algorithm overview. Given $C \times H \times W$ images from source and target domains, we pass them through the density map estimation network to obtain output predictions. A density loss is computed for source predictions based on the ground truth. In order to improve target predictions, a discriminator is used to locally classify whether a density map belongs to the source or target domain. Then, an adversarial loss is computed on the target prediction and is back-propagated to the density map estimation and counting network.

where the subscript h,w denotes a pixel in P. This loss makes the distribution of $D^{\mathcal{T}}$ closer to $D^{\mathcal{S}}$ by forcing Ψ to fool the discriminator, through the maximization of the probability of $D^{\mathcal{T}}$ being locally classified as belonging to the source domain.

Network Θ Training. Given an image I and the corresponding predicted density map D, we feed D as input to the fully-convolutional discriminator Θ to obtain the probability map P. The discriminator is trained by comparing P with the ground truth label map $Y \in \{0,1\}^{H' \times W'}$ using a pixel-wise binary cross-entropy loss

$$\mathcal{L}_{disc}(I) = -\sum_{h,w}(1 - Y_{h,w})\log(1 - P_{h,w}) + \\ +Y_{h,w} log(P_{h,w}), \tag{3}$$

where $Y_{h,w} = 0 \;\; \forall\; h,w$ if I is taken from the target domain and $Y_{h,w} = 1$ otherwise.

5 EXPERIMENTAL RESULTS

5.1 Implementation Details

Density Map Estimation and Counting Network. We build our density map estimation network based on the Congested Scene Recognition Network (CSRNet) (Li et al., 2018). Here we briefly review some of the features characterizing this algorithm. CSRNet provides a CNN-based method that can understand highly congested scenes and perform accurate density estimation and counting. It is composed of two major components. The authors use the well-known VGG-16 network (Simonyan and Zisserman, 2014) as the front-end for 2D feature extraction because of its strong transfer learning ability. On the other hand, the back-end consists of dilated kernels. The basic concept of using dilated convolutions is to deliver larger reception fields replacing the pooling operations. It is worth noting that the max pool operation is responsible for losing quality in the density generation procedure. Since the output size from VGG is reduced by a factor of 8 of the original input size, we up-sampled the final output to compare it with the ground truth density map.

Discriminator. We use a Fully Convolutional Network similar to (Tsai et al., 2018) and to (Radford et al., 2015), composed of 5 convolution layers with kernel 4×4 and stride of 2. The number of channels are $\{64,\ 128,\ 256,\ 512,\ 1\}$, respectively. Each convolution layer is followed by a leaky ReLU having a parameter equals to 0.2.

We implement the whole system using the PyTorch framework on a single Nvidia RTX 2080 GPU with 12 GB memory. To train the density estimator network and the discriminator, we use Adam optimizer (Kingma and Ba, 2014) with an initial learning rate set to 10^{-5}. During the training, it is crucial to balance the weight between density and adversarial losses. A small value of λ_{adv} may not help the training process significantly. In contrast, a larger value may propagate incorrect gradients to the density estimator. We empirically choose the value of λ_{adv} depending on the employed dataset.

5.2 Results and Discussion

We validate the proposed UDA method for density estimation and counting of traffic scenes under different settings. First, we employ the *NDISPark* dataset, and we test the *Day2Night* domain shift; then, we utilize the *WebCamT* and the *TRANCOS* datasets to take into account the *Camera2Camera* performance gap. Finally, we use the *GTA* dataset to consider the *Synthetic2Real* domain difference. For all the experiments, we base the evaluation of the models on three metrics widely used for the counting task: (i) Mean Absolute Error (MAE) that measures the absolute count error of each image; (ii) Mean Squared Error (MSE) that instead quantifies the squared count error for each image; (iii) Average Relative Error (ARE), which measures the absolute count error divided by the true count. Note that, as a result of the squaring of each error, the MSE effectively penalizes large errors more heavily than the small ones. Instead, the ARE is the only metric that considers the relation of the error and the total number of vehicles present for each image. Results are summarized in Table 1, while in the next sections, we describe the results obtained for every considered scenario. Finally, we also plot some examples of the outputs obtained using our models, showing their visual quality. In particular, Figure 5 shows the ground truth and the predicted density maps for some random samples of the considered scenarios.

5.2.1 Day2Night Domain Shift

In this scenario, we split the *NDISPark* dataset into train, validation, and test subsets containing about 100, 50, and 100 images. The former has only pictures taken during the day (source domain), while the validation and the test subsets contain night images (target domain). To fairly evaluate our method, we first consider the baseline model without the domain adaptation module (i.e., putting the λ_{adv} value to zero). Then, we add the adversarial module comparing the results. In both cases, we train the network for 300 epochs, validating at each iteration. We choose the best validation model in terms of MAE, and we test it against the test set. As showed in Table 1, using our solution, we obtained performance improvements considering all the three metrics.

5.2.2 Camera2Camera Domain Shift

In this case, we perform two sets of experiments to test the domain shift that takes place when we consider a camera different from the ones used in the training phase.

First, we consider the *WebCamT* dataset, and we split it into train, validation, and test subsets. In the former, we account for about 25,000 images belonging to 7 cameras (source domain). In the last two, we consider the remaining 15,000 pictures of 3 different cameras, having diverse contexts and slightly different angle of views (target domain). We compare the baseline and our solution when training for 20 epochs, validate it at each iteration, and choose the best model in terms of MAE.

Second, we take into account the *TRANCOS* dataset. We split it into train, validation, and test sets, following (Guerrero-Gómez-Olmedo et al., 2015). The train set represents the source domain, while the other two belong to the target domain and are collected in different contexts. We train our domain adaptation for 200 epochs, picking the best validation model in terms of MAE, and we evaluate it against the test set. We compare the obtained results with the ones claimed by (Li et al., 2018) using only the state-of-the-art CSRNet algorithm (i.e., our baseline) and with other state-of-the-art techniques present in the literature.

As showed in Table 1, we obtained performance improvements in both cases, taking into account all three metrics. Considering the publicly available *TRANCOS* dataset, we achieved superior results not only concerning the baseline but also compared to the other considered approaches.

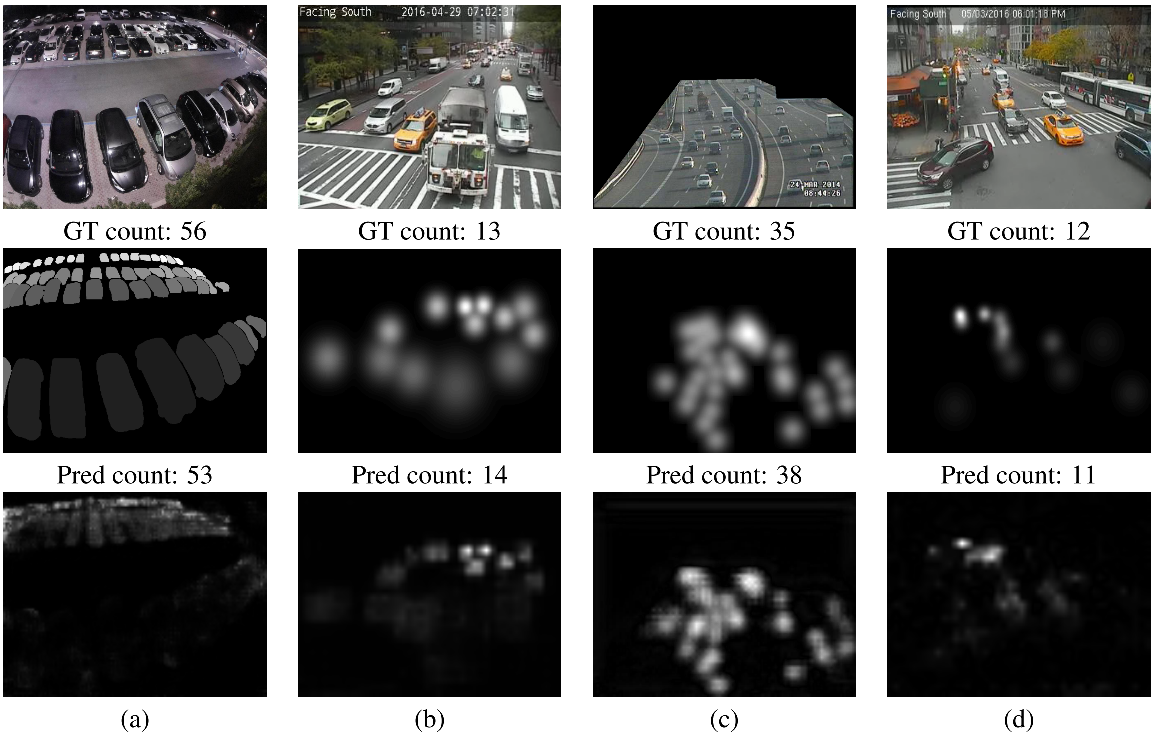
5.2.3 Synthetic2Real Domain Shift

In this scenario, we train the algorithm using synthetic images. Then we test it on real data. In particular, we consider a subset of the *GTA* dataset containing about 5,000 images of city traffic scenarios, and we use it as the training set (source domain). On the other hand, we account for the test and the validation subsets of the *WebCamT* dataset as the target domain. We compare the results obtained using the baseline model and our solution with the domain adaptation module. In both cases, we train the algorithm for 20 epochs, validating at each iteration. We choose the best model in terms of MAE.

Again, as showed in Table 1, we achieved better results compared to the basic model. We believe that

	MAE	MSE	ARE
Day2Night Domain Shift - NDISPark Dataset			
Baseline - CSRNet (Li et al., 2018)	3.95	27.45	0.43
Our Approach	**3.49**	**20.90**	**0.39**
Camera2Camera Domain Shift - WebCamT Dataset (Zhang et al., 2017a)			
Baseline - CSRNet (Li et al., 2018)	3.24	16.83	0.21
Our Approach	**2.86**	**13.03**	**0.19**
Camera2Camera Domain Shift - TRANCOS Dataset (Guerrero-Gómez-Olmedo et al., 2015)			
Hydra-CNN (Oñoro-Rubio and López-Sastre, 2016)	10.99	68.70	0.71
FCN-MT (Zhang et al., 2017a)	5.31	-	0.85
LC-ResFCN (Laradji et al., 2018)	3.32	-	-
Baseline - CSRNet (Li et al., 2018)	3.56	30.64	0.10
Our Approach	**3.30**	**23.60**	**0.08**
Synthetic2Real Domain Shift - GTA Dataset			
Baseline - CSRNet (Li et al., 2018)	4.10	25.83	0.28
Our Approach	**3.88**	**23.80**	**0.27**

Table 1: Experimental results obtained for the four considered domain shift. We employed three evaluation metrics: the Mean Absolute Error (MAE), the Mean Squared Error (MSE) and the Average Relative Error (ARE). We achieved performance improvements for all the scenarios, considering all the three metrics.


Facing South 2016-04-29 07:02:31
24 MAR-2014 08:44:26
Facing South 05/03/2016 06:01:18 PM

GT count: 56 GT count: 13 GT count: 35 GT count: 12

Pred count: 53 Pred count: 14 Pred count: 38 Pred count: 11

(a) (b) (c) (d)

Figure 5: Examples of the predicted density maps in the considered scenarios: (a) *Day2Nigh* Domain Shift using the *NDISPark* dataset; (b) and (c) *Camera2Camera* Domain Shift employing the *WebCamT* and *TRANCOS* datasets, respectively; (d) *Synthetic2Real* Domain Shift using the *GTA* dataset for the training phase and the *WebCamT* dataset for testing on real images. In the first row, we report the input images. In the second row, the ground truth, while in the third, the predicted density maps obtained with our models.

this scenario is particularly interesting because we obtained comparable results with the previous one, but this time *without* using manual annotations neither in the source domain nor in the target one.

6 CONCLUSIONS

In this article, we tackle the problem of determining the density and the number of objects present in large sets of images. Building on a CNN-based density estimator, the proposed methodology can generalize to new data sources for which there are no annotations available. We achieve this generalization by exploiting an Unsupervised Domain Adaptation strategy, whereby a discriminator attached to the output forces similar density distribution in the target and source domains. Experiments show a significant improvement relative to the performance of the model without domain adaptation. To the best of our knowledge, we are the first to introduce a UDA scheme for counting to reduce the gap between the source and the target domain without using additional labels. Given the conventional structure of the estimator, the improvement obtained by just monitoring the output entails a great capacity to generalize learned knowledge, thus suggesting the application of similar principles to the inner layers of the network.

Another contribution is represented by the creation of two new per-pixel annotated datasets made available to the scientific community. One of the two novel datasets is a synthetic dataset created from a photo-realistic video game. Here the labels are automatically assigned while interacting with the API of the graphical engine. Using this synthetic dataset, we demonstrated that it is possible to train a model with a precisely annotated and automatically generated synthetic dataset and perform UDA toward a real-world scenario, obtaining very good performance *without* using additional manual annotations.

In our view, this work's outcome opens new perspectives to deal with the scalability of learning methods for large physical systems with scarce supervisory resources.

ACKNOWLEDGEMENTS

This work was partially supported by H2020 project AI4EU under GA 825619 and by H2020 project AI4media under GA 951911.

REFERENCES

Aich, S. and Stavness, I. (2018). Improving object counting with heatmap regulation. *arXiv preprint arXiv:1803.05494*.

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

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

Amato, G., Ciampi, L., Falchi, F., Gennaro, C., and Messina, N. (2019). Learning pedestrian detection from virtual worlds. In Ricci, E., Rota Bulò, S., Snoek, C., Lanz, O., Messelodi, S., and Sebe, N., editors, *Image Analysis and Processing – ICIAP 2019*, pages 302–312, Cham. Springer International Publishing.

Boominathan, L., Kruthiventi, S. S. S., and Babu, R. V. (2016). Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 640–644, New York, NY, USA. Association for Computing Machinery.

Chen, Y., Li, W., Chen, X., and Gool, L. V. (2019). Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1841–1850.

Ciampi, L., Amato, G., Falchi, F., Gennaro, C., and Rabitti, F. (2018). Counting vehicles with cameras. In *SEBD*.

Ciampi, L., Messina, N., Falchi, F., Gennaro, C., and Amato, G. (2020a). Virtual to real adaptation of pedestrian detectors. *Sensors*, 20(18):5250.

Ciampi, L., Santiago, C., Costeira, J. P., Gennaro, C., and Amato, G. (2020b). Unsupervised vehicle counting via multiple camera domain adaptation. In Saffiotti, A., Serafini, L., and Lukowicz, P., editors, *Proceedings of the First International Workshop on New Foundations for Human-Centered AI (NeHuAI) co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago de Compostella, Spain, September 4, 2020*, volume 2659 of *CEUR Workshop Proceedings*, pages 82–85. CEUR-WS.org.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.

Guerrero-Gómez-Olmedo, R., Torre-Jiménez, B., López-Sastre, R., Maldonado-Bascón, S., and Oñoro-Rubio, D. (2015). Extremely overlapping vehicle counting. In Paredes, R., Cardoso, J. S., and Pardo, X. M., editors, *Pattern Recognition and Image Analysis*, pages 423–431, Cham. Springer International Publishing.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Hoffman, J., Wang, D., Yu, F., and Darrell, T. (2016). Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*.

Hong, W., Wang, Z., Yang, M., and Yuan, J. (2018). Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Laradji, I. H., Rostamzadeh, N., Pinheiro, P. O., Vazquez, D., and Schmidt, M. (2018). Where are the blobs: Counting by localization with point supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 547–562.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lempitsky, V. and Zisserman, A. (2010). Learning to count objects in images. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 1324–1332. Curran Associates, Inc.

Li, Y., Zhang, X., and Chen, D. (2018). Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100.

Oñoro-Rubio, D. and López-Sastre, R. J. (2016). Towards perspective-free object counting with deep learning. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 615–629, Cham. Springer International Publishing.

Radford et al., A. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. (2016). 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*, pages 3234–3243.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sindagi, V. A. and Patel, V. M. (2018). A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*, 107:3 – 16. Video Surveillance-oriented Biometrics.

Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528.

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

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176.

Zhang, S., Wu, G., Costeira, J. P., and Moura, J. M. (2017a). Understanding traffic density from large-scale web camera data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5898–5907.

Zhang, Y., David, P., and Gong, B. (2017b). Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2030.

Zhang, Y., Zhou, D., Chen, S., Gao, S., and Ma, Y. (2016). Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597.