Vehicle Crowd Analysis via Transfer Learning

Yusuf K. Hanoglu
Electronics and Comm. Engineering
Istanbul Technical University
İstanbul, Turkey

Bilge Gunsel
Electronics and Comm. Engineering
Istanbul Technical University
İstanbul, Turkey

Meltem Gulbas
Electronics and Comm. Engineering
Istanbul Technical University
İstanbul, Turkey

***Abstract*—We propose a deep learning based approach to vehicle density estimation that adopts CSRNet, originally designed for person crowd analysis, to vehicle crowd analysis. The objective is to exploit the transfer learning to accurately estimate the vehicle density with an increased learning speed. Specifically, the CSRNet architecture pre-trained on the person domain is fine tuned on the vehicle domain by feature tranformation. This is achieved by end-to-end retraining the network to output the spatial distribution of vehicles in congested scenes. The approach is evaluated on Waymo and TRANCOS data sets while ShanghaiTech data set is used for pretraining. Performance reported by the metrics of MAE and RMSE, and PSNR on different test cases, demonstrate the transfer learning significantly improves vehicle density estimation accuracy, compared to the learning from stretch. In particular, the learning accuracy achieved on Waymo, with a small size training data, is validating the potential of the approach in enhancing vehicle crowd analysis for autonomous driving task.**

***Index Terms*—Crowd analysis, vehicle density estimation, transfer learning, deep learning.**

I. Introduction

Crowd analysis aims to accurately estimate the location and count of the objects included in a congested scene. Most of the existing works focus on person crowd analysis [1]–[3] whereas some recent works paid attention to vehicles [4]. Vehicle crowd analysis is a crucial task in several applications, including traffic monitoring, event analysis and autonomous driving. Although person crowd detection and counting have made tremendous progress with the development of deep networks, the vehicle crowd analysis is still a challenging task, due to severe occlusion, scale changes, viewpoint changes, diverse visual appearances of vehicles and illumination variations.

In the literature, most of the earlier methods employ object detection or regression based approaches that fail under occlusion and scale variations [5]. Lately density-based methods that learn a spatial distribution, linearly or nonlinearly mapping the regional features to corresponding density maps, became popular because of their enhanced accuracy [2], [3], [5]. In this paper, we propose a vehicle density estimation based crowd analysis method implemented by adapting Congested Scene Recognition Network (CSRNet) [2], a deep network originally designed for person crowd analysis, to the domain of vehicle crowd analysis and density estimation. The primary objective of our work is to exploit transfer learning [6] for boosted vehicle crowd detection.

Our contributions are two folds:

- A person density estimation network is adopted to vehicle density estimation by transfer learning. It is also shown that the same network architecture can be used for learning transform from vehicle to vehicle data sets without changing the domain. Extensive evaluations on Waymo [7] and TRANCOS [4] data sets demonstrate significant improvements in accuracy, validating the potential of our method in vehicle crowd analysis.

- We have specified efficient training and testing subsets from an open access autonomous driving database, Waymo, to evaluate the learning capability of the implemented network. The size of the Waymo data set is much larger than its counterparts, making it a challenging data set to work with. Although it is employed by several object tracking [8] and segmentation [9] models for performance evaluation, as of our knowledge, this work is the first attempt on using Waymo for crowd analysis. We believe in that our training set with annotated ground truth density maps is valuable for vehicle crowd analysis research. Trained models are available on https://github.com/msprITU/Waymo-Crowd-Analysis for the usage of other researchers.

(a) (b)

Fig. 1: (a) Original image with overlaid centers. (b) Corresponding ground truth vehicle density map.

II. Distance Coding by Density Maps

The data sets prepared for training crowd detection networks consist of images depicting highly crowded vehicle coordinates (Fig. 1(a)). Corresponding ground truth (GT) density maps, coding the distance between the vehicles, are also fed into the network (Fig. 1(b)). In order to create the GT density map, the annotated vehicle center points undergo a Gaussian filtering process where the standard deviation (sigma) of the filter is determined proportionally to the average distances between vehicles and their neighbors.

Let $[n_{1_i}, n_{2_i}]$ denote the center pixel coordinate of i^{th} vehicle. Each vehicle in a 2D image plane can be modeled by a shifted dirac delta function where the amount of

shift is specified by the center pixel of the vehicle. Thus far the image of center pixels can be formulated as $\boldsymbol{C} = \sum_{i=1}^{N} \delta[n_1 - n_{1_i}, n_2 - n_{2_i}]$ where the number of vehicles included in the image is N. For an input image $\boldsymbol{a}$ sized MxL, the corresponding density map is also an MxL image $\boldsymbol{D_a}$ obtained by filtering, $\boldsymbol{D_a} = \boldsymbol{C} * * \boldsymbol{G_{\sigma_i}}$, where $\boldsymbol{G_{\sigma_i}}$ is the 2D impulse response of a Gaussian filter. Thus each point of the density map can be attained by Eq. 1.

$$d[n_1, n_2] = \sum_{k_1=0}^{L-1} \sum_{k_2=0}^{M-1} C[k_1, k_2] G_{\sigma_i}[n_1 - k_1, n_2 - k_2] \quad (1)$$

Consequently, the density map $\boldsymbol{D_a}$ is formulated as the sum of shifted Gaussian functions as in Eq. 2,

$$\boldsymbol{D_a} = \sum_{i=1}^{N} G_{\sigma_i}[n_1 - n_{1_i}, n_2 - n_{2_i}] \quad (2)$$

where Eq. 3 formulates a shifted Gaussian as a function of Euclidean distance from the center of i^{th} vehicle, $l_i = (n_1 - n_{1_i})^2 + (n_2 - n_{2_i})^2$.

$$G_{\sigma_i}[n_1 - n_{1_i}, n_2 - n_{2_i}] = \frac{1}{2\pi\sigma_i^2} exp[-\frac{l_i}{2\sigma_i^2}] \quad (3)$$

The Gaussian spread parameter σ_i is adaptively adjusted to being proportional with the distance to k-nearest neighbor vehicle centers. In congested vehicle regions corresponding Gaussian filters exhibit a smaller sigma value, indicating a higher concentration. Hence, the filter outputs at the neighboring vehicles overlap to a greater extent that can be interpreted as distance coding. This approach helps capture the varying densities and overlapping patterns of vehicles, providing a more comprehensive representation for training vehicle crowd detection networks.

III. Vehicle Density Estimation by CSRNet

We have used Congested Scene Recognition Network (CSRNet) [2] as our baseline in vehicle density estimation. CSRNet introduced for analysis of person crowds is adopted to vehicle density estimation by transfer learning. In particular the head layer of CSRNet is changed and different pre-training and fine tuning schemes are performed on the modified architecture. Note that integration of the proposed approach to the other person crowd analysis architectures is straightforward.

A. Network Architecture

CSRNet architecture listed at Table I consists of front-end and back-end modules. First 10 layers of VGG-16 that is trained with ImageNet [10] is used on front-end to obtain a pre-trained model. Back-end module is a second CNN that includes a number of dilated convolution layers (DCLs). Dilated convolution kernels skip pixels according to the dilation factor therefore cover a larger area without increasing the computational cost. Additionally DCLs eliminate the necessity of pooling layers that yields a lowered memory requirement. In this work the dilation factor is set to 2.

TABLE I: CSRNet Architecture.

input	max-pooling	conv3-512-2
front-end 10 layers of VGG-16	conv3-256-1	conv3-512-2
conv3-64-1	conv3-256-1	conv3-512-2
conv3-64-1	conv3-256-1	conv3-256-2
max-pooling	max-pooling	conv3-128-1
conv3-128-1	conv3-512-1	conv3-64-2
conv3-128-1	conv3-512-1	conv1-1-1
conv3-128-1	conv3-512-1	output
conv3-128-1	back-end	

B. Training CSRNet from Stretch

Training of the network has been performed by applying stochastic gradient descent (SGD). CSRNet architecture is trained end-to-end by starting with random weights generated from a zero mean Gaussian distribution with standard deviation 0.01.Through training all the weights are updated to be capable of extracting relevant features from the input crowd image for reconstructing the desired output density map.

To understand the whole training process, one should start with updating the final layer of the network. Let, at epoch t, the second last layer outcome of the network for a^{th} image be the feature vector $\boldsymbol{h_a^{tT}} = [h_{a_1}^t h_{a_2}^t ... h_{a_i}^t ... h_{a_P}^t]$ and the outcome of the last layer (fully connected layer) is the feature vector $\boldsymbol{z_a^t} = [z_{a_1}^t z_{a_2}^t ... z_{a_j}^t ... z_{a_R}^t]$. Thus the output of the last layer is attained by multiplying $\boldsymbol{w^t}$ with $\boldsymbol{h_{a^tT}}$, as $\boldsymbol{z_a^t} = \boldsymbol{w^t} \cdot \boldsymbol{h_a^{tT}}$, where $\boldsymbol{w^t}$ is a PxR weight matrix of the last fully connected layer. Finally the outcome of the whole network be the estimated density map $\boldsymbol{D}(\boldsymbol{h_a^t}, \boldsymbol{w^t}) = \boldsymbol{y_a^t} = [y_{a_1}^t y_{a_2}^t ... y_{a_j}^t ... y_{a_R}^t]$ obtained by applying the activation function rectified linear unit (ReLU) on $\boldsymbol{z_a^t}$ as shown in Eq. 4.

$$\boldsymbol{D}(\boldsymbol{h_a^t}, \boldsymbol{w^t}) = \text{ReLU}(\boldsymbol{z_a^t}) \quad (4)$$

As the ground truth density map being already given as $\boldsymbol{D}_a = \boldsymbol{y_a} = [y_{a_1} y_{a_2} ... y_{a_j} ... y_{a_R}]$, the loss function, which is the Euclidean distance between ground truth density map and the estimated density map, can be calculated for the minibatch which has total number of T images as in Eq. 5.

$$L(\boldsymbol{D}(\boldsymbol{h_a^t}, \boldsymbol{w^t}), \boldsymbol{D}a) = \frac{1}{2T} \sum_{a=1}^{T} ||\boldsymbol{D}(\boldsymbol{h_a^t}, \boldsymbol{w^t})^t - \boldsymbol{D}_a||_2^2 \quad (5)$$

After calculating the loss in the forward pass, the weight matrix $\boldsymbol{w}^t$ is updated by SGD in the backward pass. Let w_{ij}^t be the weight connecting $h_{a_i}^t$ and $z_{a_j}^t$. The updated weight w_{ij}^{t+1} can be calculated by Eq.6 where η is the learning rate.

$$w_{ij}^{t+1} = w_{ij}^t - \eta \frac{\partial L^t}{\partial w_{ij}^t} \quad (6)$$

Gradient of the loss shown in Eq.6 can be found by chain rule as in Eq. 7.

$$\frac{\partial L^t}{\partial w_{ij}^t} = \frac{\partial L^t}{\partial y_{a_j}^t} \frac{\partial y_{a_j}^t}{\partial z_{a_j}^t} \frac{\partial z_{a_j}^t}{\partial w_{ij}^t} \quad (7)$$

For each term on Eq. 7 taking partial derivative gives Eq. 8.

$$\frac{\partial L^t}{\partial w_{ij}^t} = (y_{a_j}^t - y_j) ReLU(y_{a_j}^t) \cdot h_{a_i}^t \quad (8)$$

C. Training by Transfer Learning

Domain adaptation of CSRNet from person to vehicle is achieved by transfer learning which aims to improve learning efficiency by adapting learned weights of a pre-trained model to the target model [6]. In our work, the network trained from stretch on ShanghaiTech A/B data set is used as the pre-trained model for end-to-end training of CSRNet on TRANCOS as well as Waymo data sets. This scheme is referred as learning transform from person to vehicle. Additionally, CSRNet is trained from stretch by using TRANCOS data set. The fine tuning has been performed on Waymo data set that leads learning transform from vehicle to vehicle. Moreover we trained CSRNet on Waymo from strecth. Fig. 2 illustrates loss versus epoch number for our training from stretch by Waymo compared to the loss of the transfer learning based training schemes. To highlight the difference, the first 30 epoch is reported. As seen in Fig. 2, the transfer learning significantly increases the convergence speed so as the learning capability of the network.

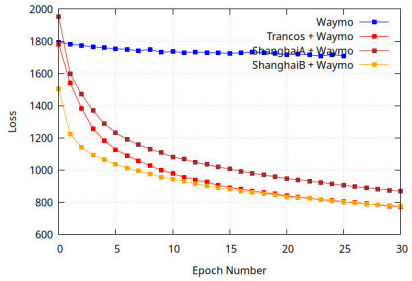


Fig. 2: Change of loss through training.

IV. Performance Evaluation

A. Test Data

ShanghaiTech data set [1] consists of people crowd images collected from internet and streets of Shanghai. ShanghaiTech $Part\_A$ has total of 482 images with 241,677 annotated heads. Average number of people is 501.2 per image. 300 of the images are used to training while 182 images are separated to be used for testing. ShanghaiTech $Part\_B$ has 716 images with 123.6 average head count. Train set is composed of 400 images which leaves 316 images for the test set.

TRANCOS data set [4] is created for overlapping vehicle counting from traffic surveillance cameras in Spain. There are a total of 13,543 annotated vehicles on the 403 training data set images. Test set consists of 420 images with a total of 16,666 annotated vehicles.

Waymo data set [7] is a comprehensive collection of sensor data captured from Waymo's self-driving vehicles during real-world driving scenarios. It includes video data simultaneously captured by 5 high-resolution CCD cameras, and lidar-radar sensors, providing a rich and detailed representation of the environment surrounding the vehicle. In our tests we used 132 of the front CCD camera sequences where the training set consists of 42 video sequences with a total of 9935 annotated vehicles. Test set includes 90 video sequences having 309965 annotated vehicles.

B. Performance Metrics

While loss is related to how accurate the density map is estimated, the peak signal-to-noise-ratio (PSNR) in dB, formulated in Eq. 9, is used for performance evaluation,

$$PSNR = 10log_{10}\left(\frac{255^2}{\frac{1}{T}\sum_a^T|\hat{\boldsymbol{D}}_a - \boldsymbol{D}_a|^2}\right) \tag{9}$$

where $\hat{\boldsymbol{D}}_a$ and $\boldsymbol{D}_a$ denote the estimated and ground truth vehicle density map, respectively.

We have also employed commonly used mean absolute error (MAE) and the root of MAE (RMSE) metrics in evaluation. These metrics dependent on how proximate the total count of vehicle is predicted. Notating the difference between the estimated count as v_a and the actual count of vehicles on the images as $\hat{v_a}$, Eq.10 formulates MAE and RMSE for T images. Knowing that each density map is modeled by a mixture of Gaussians (Eq. 2), the estimated vehicle count ($\hat{v_a}$) can be calculated by taking the integral of the estimated density map.

$$MAE = \frac{1}{T}\sum_i^N|v_a - \hat{v_a}|, RMSE = \sqrt{\frac{1}{T}\sum_{a=1}^T|v_a - \hat{v_a}|^2} \tag{10}$$

C. Evaluation Test Cases

Test Case 1 is designed for evaluation of improvement attained from transfer learning . First CSRNet is trained from stretch on Waymo as well as TRANCOS data set. MAE, RMSE and PSNR values are reported on Waymo and TRANCOS test sets (yellow row on Table II and Table III). For transfer learning, CSRNet pretrained on ShanghaiTech A data set from stretch is tuned on Waymo. The test is repeated by using TRANCOS for fine tuning in the transfer learning phase. Results demonstrate that, the feature learning transfer from person to vehicle significantly reduced MAE and RMSE in both the Waymo and TRANCOS test sets. Also PSNR is highly improved especially for Waymo Test Set (green rows on Table II and Table III). In order to see the impact of learning transfer from a vehicle domain to another vehicle domain, CSRNet pre-trained on TRANCOS is fine tuned on Waymo data set. Among the conducted transfer learning tests, this case exhibited the best performance on the Waymo test set for all metrics (red row on Table II and Fig. 3).

Test Case 2 is designed to report the performance achieved on TRANCOS compared to the baseline works. MAE achieved by our model without transfer learning is very close to MAE reported in [4] (orange and yellow rows on Table III). However our model highly decreases the MAE (11.05) via transfer learning attained by fine tuning on pre-trained Shanghai Tech Part A (red row on Table III). We have also evaluated the impact of transfer learning by comparing the results with [2] where the training and evaluation are performed only on the masked vehicle regions of the TRANCOS images with a fixed σ_i and the training set is extended by validation set

TABLE II: Impact of transfer learning on Waymo.

Training	Pretraining	Waymo Test Set		
		MAE↓	RMSE↓	PSNR↑
Waymo	-	11.69	12.63	27.66
Waymo	ShangT A	8.32	8.99	31.27
Waymo	TRANCOS	7.26	7.95	31.87

TABLE III: Impact of transfer learning on TRANCOS compared to the existing models.

Training	Pretraining	TRANCOS Test Set		
		MAE↓	RMSE↓	PSNR↑
TRANCOS [4]	-	17.68	-	-
TRANCOS	-	17.45	20.50	21.90
TRANCOS	ShangT A	11.05	15.11	24.55
MTRANCOS [2]	-	3.56	-	27.10
MTRANCOS	-	3.46	5.00	34.15
MTRANCOS	ShangT A	3.37	4.71	34.65

of TRANCOS. Results reported on Table III as MTRANCOS demonstrate that the enlarged training set with masked region of interest significantly reduces MAE and improves PSNR (blue row). Our training provides a slightly better performance compared to [2] (purple row). Transfer learning slightly improves the performance (green row) .

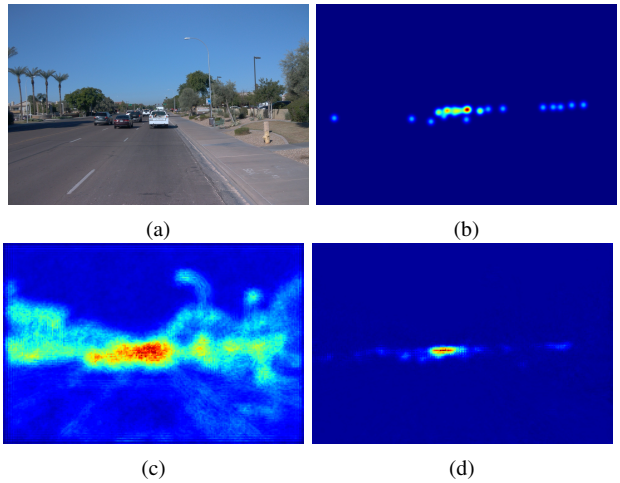
Test Case 3 deals with evaluation of the crowd detection accuracy for different sizes of vehicles. CSRNet models trained for Test Case 1 are employed for the evaluation. MAE values are reported for different sizes of vehicles: large (L), medium (M), and small (S). The vehicle sizes specified by k-means clustering of GT bounding box sizes on Waymo are: S: 34,384 pixels, M: 190,233 pixels, and L: 381,086 pixels. The number of vehicles in each group is as follows: S: 11,337, M: 1,717, and L: 4,431. As it is reported at Table IV, across all training instances conducted with Waymo, the highest performance is achieved on small vehicles, however accuracy attained at all sizes are not significantly different. This implies that, with the help of density maps, CSRNet enables to learn vehicle detection in crowded regions. Moreover accuracy on Waymo test set has been significantly increased as a result of the transfer learning. Its positive impact on the performance of vehicles classified as very small is greater.

V. Conclusions

We propose a transfer learning based density estimation approach for vehicle crowd analysis. It can be integrated with any of the existing person crowd density estimation network as it is. Our notable contribution lies in successfully utilizing the challenging Waymo data set, sourced from an open-access autonomous driving database, to evaluate our network's learning capabilities.

TABLE IV: Accuracy attained for different vehicle sizes.

Transfer L	Pretraining	Waymo Test Set (MAE↓)		
		Small	Medium	Large
Waymo	-	10.78	13.66	12.82
Waymo	ShangT A	6.80	10.03	9.11
Waymo	TRANCOS	6.33	9.19	8.51



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

Fig. 3: (a) A video frame from Waymo segment 1179. (b) Corresponding ground truth vehicle density map. Density map estimated c) by only Waymo training and d) by transfer learning with pretraining on TRANCOS.

Through extensive experiments and evaluations, we demonstrate the efficacy of the transfer learning in vehicle crowd analysis. It is shown that the transfer learning highly increases the learning speed while reducing the size of annotated training data significantly. In our case, we used only 9935 video frames sampled from 42 sequences for tuning on Waymo, where the Waymo test set includes 309965 frames. Also, as a result of 55 epoches pre-training on Shangai Tech data, transfer from person to vehicle domain takes only 100 and 60 epoches on Waymo and Transcos, respectively. Our results also demonstrate notable improvement in the representation quality of the estimated density maps, validating the potential of our approach in the context of autonomous driving.

References

[1] Z. et al., "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 589–597.

[2] Y. Li, X. Zhang, and D. Chen, "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.

[3] C. et al., "Rethinking spatial invariance of convolutional networks for object counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 638–19 648.

[4] O. et al., "Extremely overlapping vehicle counting," in *Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, 2015.

[5] M. A. K. et al., "Revisiting crowd counting: State-of-the-art, trends, and future perspectives ," *Image and Vision Computing*, vol. 129, 2023.

[6] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. on Knowledge and Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[7] S. et al., "Scalability in perception for autonomous driving: Waymo open dataset," 2020.

[8] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 11 784–11 793.

[9] A. Z. Zhu, V. Casser, R. Mahjourian, H. Kretzschmar, and S. Pirk, "Instance segmentation with cross-modal consistency," 2022.

[10] R. et al., "ImageNet Large Scale Visual Recognition Challenge," *Int. Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.