

FOCUS ON SEMANTIC CONSISTENCY FOR CROSS-DOMAIN CROWD UNDERSTANDING

Tao Han, Junyu Gao, Yuan Yuan, Qi Wang*

School of Computer Science and Center for OPTical IMagery Analysis and Learning(OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P.R.China

ABSTRACT

For pixel-level crowd understanding, it is time-consuming and laborious in data collection and annotation. Some domain adaptation algorithms try to liberate it by training models with synthetic data, and the results in some recent works have proved the feasibility. However, we found that a mass of estimation errors in the background areas impede the performance of the existing methods. In this paper, we propose a domain adaptation method to eliminate it. According to the semantic consistency, a similar distribution in deep layer's features of the synthetic and real-world crowd area, we first introduce a semantic extractor to effectively distinguish crowd and background in high-level semantic information. Besides, to further enhance the adapted model, we adopt adversarial learning to align features in the semantic space. Experiments on three representative real datasets show that the proposed domain adaptation scheme achieves the state-of-the-art for cross-domain counting problems.

Index Terms— Crowd counting, domain adaptation, semantic consistency, adversarial learning

1. INTRODUCTION

Over the past few years, crowd understanding has become increasingly influential in the field of computer vision. Because it plays an important role in social management, such as video surveillance, public area planning, crowd congestion warning, traffic flow monitoring and so on [1–3]. Crowd counting is a foundation of crowd understanding, this task strives to understand people from images, predict density maps and estimate the number of pedestrians for crowd scenes. At present, many CNN-based approaches [4–8] have achieved phenomenal performance, and their success is driven by the availability of public crowd datasets. Unfortunately, the existing datasets (such as Shanghai Tech A/B [9], and UCF-QNRF [10], etc.) are so small-scale that makes it difficult for trained models to perform well in other scenarios. The high dependency on annotated data makes it difficult to deploy to social management.

To address the problem of data scarcity in crowd counting, some works [11–13] explore unsupervised or weakly supervised crowd counting, but these methods did not completely get rid of the dependence on manually annotated data. Inspired by the application of synthetic data [14, 15] in other visual fields, a large-scale crowd counting dataset named GCC was established by Wang *et al.* [16], where the data is generated and annotated automatically by a computer game mod. Although this novel data generation method solves the challenge of manually labeling data, one problem that comes with it is that the gap between the synthetic scenes and the real-world scenes is too large. Therefore, a well-trained model on the GCC dataset doesn't work well in the real world. Some of the recent work [17–21] provided us with a domain adaptation strategy, using image style transfer networks to narrow the domain gap. For the cross-domain crowd counting problem, [16] also proposed a domain adaptation method to make synthetic data closer to real data in visual perception via the SE Cycle GAN [16] network. This method of using image style transfer has achieved better results than the CNN models without domain adaptation, but the regression error in the background area reduces counting performance.

In this paper, the central issue is *how to design a better domain adaptation scheme for reducing the counting noise in the background area*. After observing a large number of images in the GCC [16] dataset, we find that the characters in the synthetic scene and the real scene have a high degree of visual similarity, while the background has a large gap. This similarity is more obvious in the high-level semantic representation. We assume that if the network can pay more attention to the semantic consistency of the crowd, it will help to narrow the domain gap. To make our adapted model can extract the semantic consistency feature for synthetic and real-world data, we first introduce a semantic extractor by further exploiting the semantic label. Considering that GCC [16] dataset provides the mask for crowd area, correspondingly, we train a segmentation model and get the semantic label for real datasets in free. Furthermore, we adopt adversarial learning to align the semantic space. Based on the above two methods, our domain adaption framework is formed. A detailed description of the network will be in the second 2.

In summary, our main contributions of this paper are as

* Qi Wang is the corresponding author. This work was supported by the National Key R&D Program of China under Grant 2017YFB1002202, National Natural Science Foundation of China under Grant U1864204, 61773316, 61632018 and 61825603.

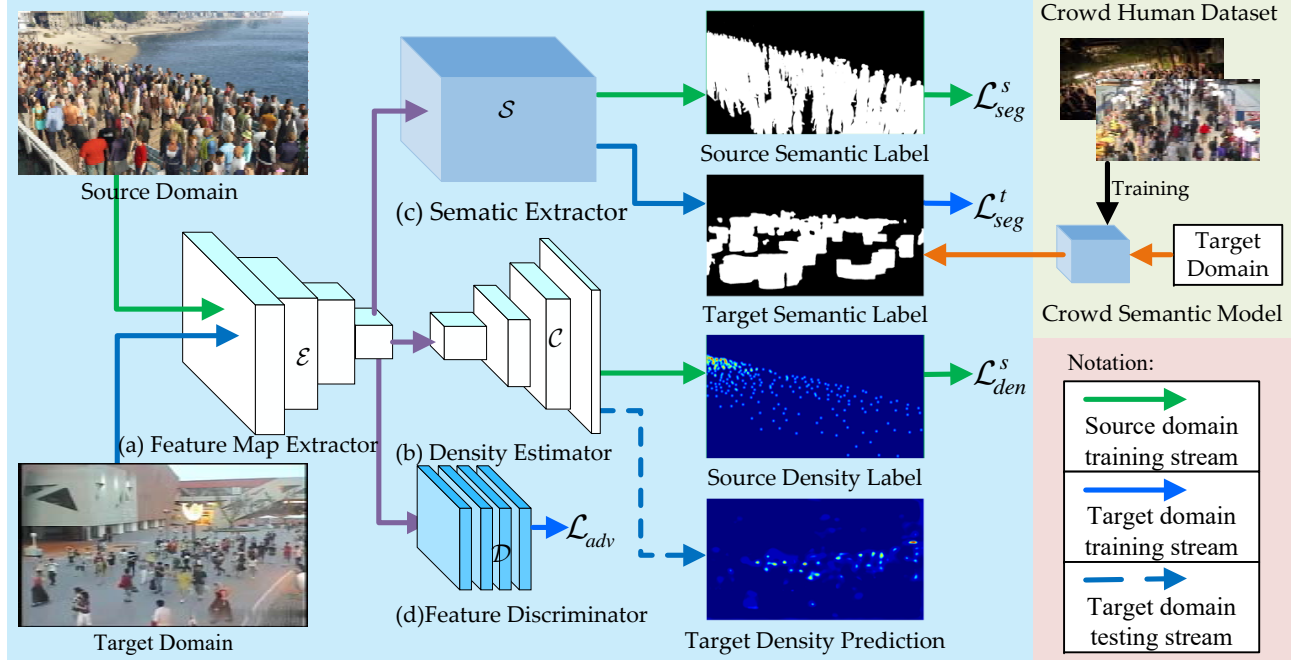


Fig. 1. The architecture of the proposed domain adaptation network for across-domain crowd understanding.

follows:

- We exploit a large-scale human detection dataset to train a crowd semantic model, which can generate crowd semantic labels for all real crowd datasets.
- We propose a domain adaptive network based on semantic consistency, which strives to focus on the consistent feature of the cross-domain crowd.
- We apply our framework to three real datasets, and it yields a new record on the across-domain crowd counting problem.

2. PROPOSED FRAMEWORK

The overall architecture of our proposed domain adaptation framework is illustrated in Fig.1. In this section, we first describe the details of the architecture, then introduce the various loss functions, and finally, we give the training details of the framework.

2.1. Framework Details

For the sake of understanding, some symbolic definitions are given here. In this paper, the number of available annotation data is N_s source domain images and N_t target domain images, denoted as $X^s = \{(x_i^s, y_i^s, z_i^s)\}_{i=1}^{N_s}$ and $X^t = \{(x_i^t, z_i^t)\}_{i=1}^{N_t}$ respectively. Each image $x_i^s, x_i^t \in \mathbb{R}^{H \times W \times 3}$ has RGB three-channel pixels with height H , width W . The source domain image x_i^s has one-channel per-pixel label of head position $y_i^s \in \{0, 1\}^{H \times W}$ and crowd mask $z_i^s \in \{0, 1\}^{H \times W}$, while the target image x_i^t only has a roughly crowd mask $z_i^t \in \{0, 1\}^{H \times W}$. As shown in Fig. 1 (a), (b),

(c) and (d), there are four sub-networks, namely, the common feature map extractor \mathcal{E} , the crowd density estimator \mathcal{C} , the crowd semantic extractor \mathcal{S} , and the feature discriminator \mathcal{D} . They are parameterized by $\theta_e, \theta_c, \theta_s$ and θ_d , respectively.

Feature Map Extractor Considering that the VGG network has less computation and strong feature representation ability, we choose a pre-trained VGG16 [22] with batch normalization as the front-end feature map extractor. It is also fair to compare with the previous work. Given a pair of images x^s and x^t as input, the output produced by the feature extractor \mathcal{E} can be represented by the following mapping:

$$f^s = \mathcal{E}(x^s; \theta_e), f^t = \mathcal{E}(x^t; \theta_e). \quad (1)$$

Semantic Extractor(SE) The semantic extractor is designed to predict the crowd mask. We propose a crowd semantic extractor by slightly modifying the pyramid module of PSPNet proposed by Zhao *et al.* [23], which fuses features under four different pyramid scales to aggregate context information for different regions. The fusion feature outputs the final semantic prediction \hat{z} through the module, which is defined as follows:

$$\hat{z}^s = \mathcal{S}(f^s; \theta_s), \hat{z}^t = \mathcal{S}(f^t; \theta_s). \quad (2)$$

Feature Discriminator(FD) The feature discriminator achieves domain adaptation by identifying whether the input features f come from the source domain or the target domain. We use an architecture similar to [24]. The network consists of 4 convolution layers with kernel 4×4 and stride of 2. Given the features f^s and f^t , we forward f^s and f^t to a feature discriminator \mathcal{D} . The discriminant result p^s, p^t can be

described by the following mapping:

$$p^s = \mathcal{D}(f^s; \theta_d), p^t = \mathcal{D}(f^t; \theta_d). \quad (3)$$

Density Estimator Since the core of this paper is not the design of crowd density predictor, we only adopt a series of simple convolution and up-sample layers to build our crowd density predictor. As shown in Fig. 1 (b), each convolution 3×3 layer is followed by a up-sample layer. Finally, the source domain density map participating in the training is defined as:

$$\hat{y}^s = \mathcal{C}(f^s; \theta_c). \quad (4)$$

2.2. Loss Functions

In section 2.1, we define the outputs $\{\hat{y}^s, \hat{y}^t\}$, $\{p^t, p^t\}$ and \hat{z}^s . Correspondingly, we define some loss functions to train our model. The training of the proposed framework is to minimize a weighted combination loss function with respect to the parameters $\{\theta_e, \theta_s, \theta_c, \theta_d\}$ of the sub-networks. The final objective function is summed as:

$$\mathcal{L}(X^s, X^t; \theta_e, \theta_s, \theta_c) = \mathcal{L}_{den}^s + \lambda^s \mathcal{L}_{seg}^s + \lambda^t \mathcal{L}_{seg}^t + \lambda^d \mathcal{L}_{adv}, \quad (5)$$

where λ^s, λ^t denote the weighting parameters of the different mask segmentation loss functions. Since the loss of mask segmentation is designed to assist the network to focus on semantic consistency of the crowd area, its weight should be set carefully. After the experimental testing, we found that it better to set them all at 0.01. λ^d is chosen empirically to strike a balance among the model capacity. In the following, we elaborate on each of these loss functions.

Crowd Density Estimation Loss To predict density maps \hat{y}^s for source-domain images x^s , the density estimation loss \mathcal{L}_{den}^s given by the typical Euclidean distance based on the source domain ground truths y^s is to supervised train the feature extractor \mathcal{E} and the crowd density predictor \mathcal{C} . In symbols, it is defined as follows:

$$\mathcal{L}_{den}^s(\theta_e, \theta_c) = \frac{1}{2N} \sum_{i=1}^N \|\hat{y}_i^s - F(y_i^s)\|^2, \quad (6)$$

where $F(y_i^s)$ is the gaussian density map of y_i^s , which is generated following a gaussian kernel function of [4]. \hat{y}_i^s is the estimated density map of x_i^s , which comes from the mapping defined in Equ. 4.

Crowd Semantic Segmentation Loss The \mathcal{L}_{seg} is designed as a auxiliary loss to extractor the semantic consistency feature for different domain's crowd. The goal of the semantic extractor is to learn correspondingly from the input characteristics f^s, f^t to predict the crowd mask z^s, z^t . To train the network, we first introduce the source domain crowd segmentation loss functions \mathcal{L}_{seg}^s , which is a binary class entropy, defined as:

$$\mathcal{L}_{seg}^s(\theta_e, \theta_s) = -\frac{1}{N} \sum_{i=1}^N (z_i^s \log \hat{z}_i^s + (1 - z_i^s) \log (1 - \hat{z}_i^s)), \quad (7)$$

where z_i^s is the source domain crowd semantic label. \hat{z}_i^s is

the probability of each pixel in the semantic prediction map activated by a sigmoid function.

As mentioned in section 1, we train a model with the CrowdHuman [25] dataset to generate the semantic label z^t for free. Since the center of this paper is not on how to acquire the semantic label, it would not be elaborated here due to the limited space. The visualization of z^t can be seen in figure 1. It can be observed that the target domain semantic label z^t is not as reliable as the source domain z^s because the pedestrian label is rectangular in object detection. To eliminate the negative effects of such inaccurate segmentation, we only use background labels to promote training, that is, we ignore the white region of the semantic lable z^t . Mathematically, we designed Equ. 8 to filter the prediction mask \hat{z}^t .

$$H(\hat{z}^t, z^t) = \hat{z}^t (1 - z^t) + z^t, \quad (8)$$

let $\bar{z}^t = H(\hat{z}^t, z^t)$, we define loss function \mathcal{L}_{seg}^t as follow:

$$\mathcal{L}_{seg}^t(\theta_e, \theta_s) = -\frac{1}{N} \sum_{i=1}^N (z_i^t \log \bar{z}_i^t + (1 - z_i^t) \log (1 - \bar{z}_i^t)). \quad (9)$$

Semantic Space Adversarial Loss In the hopes of making the \mathcal{E} extract consistent features for source domain and target domain, we introduce an adversarial loss \mathcal{L}_{adv} following [24]. Specifically, we first train a discriminator \mathcal{D} to distinguish between the source feature f^s and target feature f^t by minimizing a supervised domain loss. (i.e. \mathcal{D} should ideally output P^s to 1 in the source feature f^s and p^t to 0 for that in the target feature f^t). We then update the \mathcal{E} to fool the discriminator \mathcal{D} by inverting its output p^t from 0 to 1, that is, by minimizing

$$\mathcal{L}_{adv}(\theta_e) = -\sum_{h,w} \log(p^t), \quad (10)$$

where h and w are the coordinate dimensions of the output p^t .

2.3. Implementation

Scene Regularization GCC [16] is a large-scale synthetic dataset that includes a variety of scenarios, weather, and time periods. Therefore, adding the entire dataset to the training will bring negative effects. To eliminate this adverse effect and facilitate fair comparison, we adopted the Scene Regularization [16] to select the images.

Training Details During the training phase, the goal is to optimize \mathcal{E} , \mathcal{S} , and \mathcal{C} . Due to limited memory, we set the batch size to 8 and perform randomly cropping the images with a size of 480×640 . and we adopt the Adam algorithm to optimize the network, the initialization learning rate is set to 10^{-5} . The training and evaluation are performed on NVIDIA GTX 1080Ti GPU using the C^3 framework [27].

3. EXPERIMENTAL RESULTS

3.1. Evaluation Metrics

Count Error Following the convention of existing works [28, 29], we adopt Mean Absolute Error (MAE) and Mean

Table 1. The performance of No Adaptation, Cycle GAN, SE Cycle GAN and our approaches on the three real-world datasets.

Method	DA	ShanghaiTech Part A				ShanghaiTech Part B				UCF-QNRF			
		MAE	MSE	PSNR	SSIM	MAE	MSE	PSNR	SSIM	MAE	MSE	PSNR	SSIM
NoAdpt [16]	✗	160.0	216.5	19.01	0.359	22.8	30.6	24.66	0.715	275.5	458.5	20.12	0.554
Cycle GAN [26]	✓	143.3	204.3	19.27	0.379	25.4	39.7	24.60	0.763	257.3	400.6	20.80	0.480
SE Cycle GAN [16]	✓	123.4	193.4	18.61	0.407	19.9	28.3	24.78	0.765	230.4	384.5	21.03	0.660
NoAdpt(ours)	✗	190.8	298.1	20.57	0.457	24.6	33.7	24.14	0.687	296.1	467.9	20.46	0.513
SE+FD	✓	129.3	187.6	21.58	0.513	16.9	24.7	26.20	0.818	221.2	390.2	23.10	0.708

Squared Error (MSE) as count error evaluation metrics, which are formulated as below:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|, MSE = \sqrt{\frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|^2}, \quad (11)$$

where N is the number of testing images, y_i is the ground truth counting value and \hat{y}_i is the estimated counting value for the i th test image.

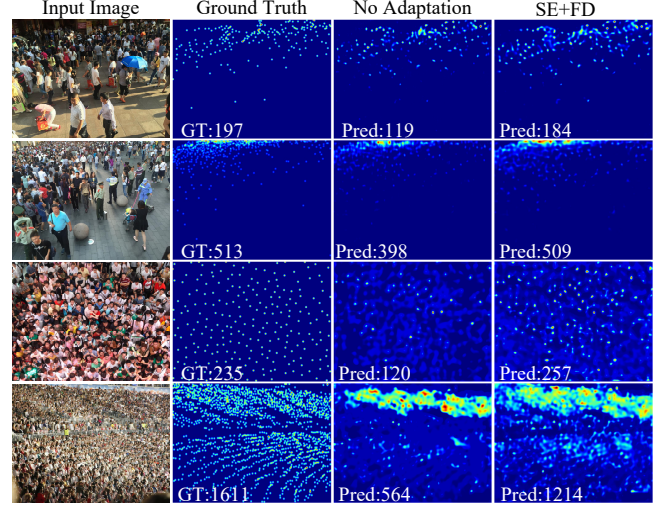
Density Map Quality Besides, to evaluate the quality of the predicted density maps, we also introduce PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity in Image)[30].

3.2. Datasets

We conduct the experiments on the ShanghaiTech PartA/B dataset [9] and UCF-QNRF dataset [10]. The ShanghaiTech Part A contains 482 crowd images (300 training and 182 testing images), and the average number of the pedestrian is 501. The ShanghaiTech Part B is with 716 images (400 training and 316 testing images), and the average number of people per image is about 123. The UCF-QNRF is a congested crowd dataset, which consists of 1,535 images(1201 training and 334 testing images), with the count ranging from 49 to 12,865, and the average number of the pedestrian is 815 per image.

3.3. Performance Comparison

In this section, we perform experiments on three typical datasets for across domain crowd counting and then compare the performance of our proposed method with the state-of-the-art SE Cycle GAN [16]. The test results are shown in Table 1. Compared with Cycle GAN [26] and SE Cycle GAN [16], which adopt image style transfer, our approach is more practical and yields better results. For Shanghai B, with no domain adaptation, the results of our baseline model are not as good as the SFCN [16], because we adopt a simpler crowd counter. However, After the domain adaptation, our model achieves comprehensive transcendence in all metrics. Our proposed method reduces the MAE to 16.9 and the MSE to 24.7, which dropped by 31.3% and 26.7% compared with the baseline model. In terms of image quality, we also

**Fig. 2.** Exemplar results of adaptation from GCC to Shanghai Tech Part B and UCF-QNRF datasets.

obtain better SSIM and PSNR. For the Shanghai B and UCS-QNRF datasets, both are the congested datasets, our proposed method also performs a better result than the state-of-the-art SE Cycle GAN[16].

Fig.2 shows the visualization results on the real datasets. It can be observed visually that the Column 4 with domain adaptation is closer to the ground truth than Column 3 without domain adaptation in terms of image quality and crowd counting. This improvement is because our method focuses on the consistency of crowd features, thus reducing the estimation error of background.

4. CONCLUSION

In this paper, we aim to count people for real scenarios using synthetic datasets. For the problem of background estimation error in the existing methods, we propose an effective domain adaptation framework, which emphasizes the network to concentrate on the crowd-area semantic consistency of the source domain and target domain by using two adaptation strategies. Experiments on high-density and low-density datasets show that our proposed method achieves state-of-the-art performance. In future work, we will further utilize high-level semantic information and domain transfer to achieve higher precision crowd counting.

References

- [1] D. B. Sam, S. Surya, and R. V. Babu, “Switching convolutional neural network for crowd counting,” in *CVPR*. IEEE, 2017, pp. 4031–4039.
- [2] Y Wang, J Hou, and L Chau, “Object counting in video surveillance using multi-scale density map regression,” in *ICASSP*. IEEE, 2019, pp. 2422–2426.
- [3] Q Wang, M Chen, F Nie, and X Li, “Detecting coherent groups in crowd scenes by multiview clustering,” *T-PAMI*, vol. 42, no. 1, pp. 46–58, 2020.
- [4] Y. Li, X. Zhang, and D. Chen, “Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes,” in *CVPR*. IEEE, 2018, pp. 1091–1100.
- [5] X. Wu, Y. Zheng, H. Ye, W. Hu, J. Yang, and L. He, “Adaptive scenario discovery for crowd counting,” in *ICASSP*. IEEE, 2019, pp. 2382–2386.
- [6] X Ding, Z Lin, F He, Y Wang, and Y Huang, “A deeply-recursive convolutional network for crowd counting,” in *ICASSP*. IEEE, 2018, pp. 1942–1946.
- [7] J. Gao, Q. Wang, and Y. Yuan, “Scar: Spatial-/channel-wise attention regression networks for crowd counting,” *Neurocomputing*, vol. 363, pp. 1–8, 2019.
- [8] J Gao, Q Wang, and X Li, “Pcc net: Perspective crowd counting via spatial convolutional network,” *IEEE Transactions on Circuits and Systems for Video Technology*, DOI: 10.1109/TCSVT.2019.2919139.
- [9] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, “Single-image crowd counting via multi-column convolutional neural network,” in *CVPR*. IEEE, 2016, pp. 589–597.
- [10] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, “Composition loss for counting, density map estimation and localization in dense crowds,” in *ECCV*, 2018, pp. 532–546.
- [11] Chen Change Loy, Shaogang Gong, and Tao Xiang, “From semi-supervised to transfer counting of crowds,” in *ICCV*, 2013, pp. 2256–2263.
- [12] X. Liu, J. van de Weijer, and A. D. Bagdanov, “Leveraging unlabeled data for crowd counting by learning to rank,” in *CVPR*, 2018, pp. 7661–7669.
- [13] D. B. Sam, N. N. Sajjan, H. Maurya, and R. V. Babu, “Almost unsupervised learning for dense crowd counting,” in *AAAI*, 2019, vol. 27.
- [14] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, “Playing for data: Ground truth from computer games,” in *ECCV*. Springer, 2016, pp. 102–118.
- [15] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, “The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” in *CVPR*, 2016, pp. 3234–3243.
- [16] Q. Wang, J. Gao, W. Lin, and Y. Yuan, “Learning from synthetic data for crowd counting in the wild,” in *CVPR*, 2019, pp. 8198–8207.
- [17] J. Hoffman, E. Tzeng, T. Park, J. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, “Cycada: Cycle-consistent adversarial domain adaptation,” in *ICML*, 2018, pp. 1994–2003.
- [18] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, “Learning from simulated and unsupervised images through adversarial training,” in *CVPR*, 2017, pp. 2107–2116.
- [19] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, “Image to image translation for domain adaptation,” in *CVPR*, 2018, pp. 4500–4509.
- [20] C Lee, T Batra, M Baig, and D Ulbricht, “Sliced wasserstein discrepancy for unsupervised domain adaptation,” in *CVPR*, June 2019.
- [21] Y Pan, T Yao, Y Li, Y Wang, C Ngo, and T Mei, “Transferable prototypical networks for unsupervised domain adaptation,” in *CVPR*, June 2019.
- [22] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [23] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *CVPR*, 2017, pp. 2881–2890.
- [24] Yi Tsai, W Hung, S Schuler, K Sohn, M Yang, and M Chandraker, “Learning to adapt structured output space for semantic segmentation,” in *CVPR*, 2018, pp. 7472–7481.
- [25] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, “Crowdhuman: A benchmark for detecting human in a crowd,” *arXiv preprint arXiv:1805.00123*, 2018.
- [26] J. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *CVPR*, 2017, pp. 2223–2232.
- [27] J Gao, W Lin, B Zhao, D Wang, Cu Gao, and J Wen, “C³ framework: An open-source pytorch code for crowd counting,” *arXiv preprint arXiv:1907.02724*, 2019.
- [28] C. Zhang, H. Li, X. Wang, and X. Yang, “Cross-scene crowd counting via deep convolutional neural networks,” in *CVPR*, 2015, pp. 833–841.
- [29] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, “Single-image crowd counting via multi-column convolutional neural network,” in *CVPR*, 2016, pp. 589–597.
- [30] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, et al., “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.