Fine-grained Domain Adaptive Crowd Counting via Point-derived Segmentation

Yongtuo Liu¹,², Dan Xu³, Sucheng Ren¹, Hanjie Wu¹, Hongmin Cai¹, Shengfeng He¹,⁴*

¹School of Computer Science, South China University of Technology
²Institute of Informatics, University of Amsterdam
³Department of Computer Science and Engineering, Hong Kong University of Science and Technology
⁴School of Computing and Information Systems, Singapore Management University

Abstract—Due to domain shift, a large performance drop is usually observed when a trained crowd counting model is deployed in the wild. While existing domain-adaptive crowd counting methods achieve promising results, they typically regard each crowd image as a whole and reduce domain discrepancies in a holistic manner, thus limiting further improvement of domain adaptation performance. To this end, we propose to untangle domain-invariant crowd and domain-specific background from crowd images and design a fine-grained domain adaption method for crowd counting. Specifically, to disentangle crowd from background, we propose to learn crowd segmentation from point-level crowd counting annotations in a weakly-supervised manner. Based on the derived segmentation, we design a crowd-aware domain adaptation mechanism consisting of two crowd-aware adaptation modules, i.e., Crowd Region Transfer (CRT) and Crowd Density Alignment (CDA). The CRT module is designed to guide crowd features transfer across domains beyond background distractions. The CDA module dedicates to regularising target-domain crowd density generation by its own crowd density distribution. Our method outperforms previous approaches consistently in the widely-used adaptation scenarios.

I. INTRODUCTION

Crowd counting has drawn increasing attention because of its fundamental role in social management [1], [2]. Due to domain shift [3], performance usually degrades a lot when trained crowd counting models are deployed in unseen crowd scenes. To fill the performance gap, a direct solution is to massively label abundant images in each crowd scene. However, the labeling is quite onerous for crowd counting as it requires labeling all human heads in each crowd image.

To avoid labeling burdensome, one promising way is to introduce Unsupervised Domain Adaptation (UDA) to transfer essential knowledge learned from a labeled source domain to a related but unlabeled target domain [5]. Recently, several methods are proposed to apply UDA for domain-adaptive crowd counting, including pixel-level adaptation methods [6], [7] and feature-level adaptation methods [8]–[11]. The feature-level methods can achieve competitive performance and work efficiently, thus dominating the existing literature.

While achieving promising results, existing domain adaptive crowd counting methods reduce domain discrepancies on crowd and background simultaneously. The holistic manner inevitably degrades domain adaptation performance considering that domain-specific background varies a lot across domains (shown in Fig. 1) and background alignment across domains challenges domain-invariant representation learning, which further harms the discrimination of crowd and background critical for crowd counting [12], [13].

To this end, we propose to treat crowd and background differently while conducting domain adaptation. Note that crowd counting only labels one point per human without segmentation. To untangle crowd and background from point-level annotations, we learn crowd segmentation from the sparse point annotations in a weakly-supervised manner. Based on the derived segmentation, we propose a Crowd-aware domain Adaptation framework for Crowd Counting (CACC), which consists of two crowd-aware adaptation modules, namely Crowd Region Transfer (CRT) and Crowd Density Alignment (CDA). Specifically, to guide crowd alignment across domains beyond background distractions, we introduce the CRT module to bridge domains by learning domain-invariant crowd features. Besides, we introduce the CDA module to generate segmentation-guided pseudo labels in the target domain to regularize crowd density generation by target-domain’s own
crowd density distribution, instead of by source-domain crowd density distribution utilized in the previous methods [9]–[11]. The design considers that different domains usually have quite different crowd density distributions, as shown in Fig. 1. It inevitably degrades the adaptation performance to directly utilize source-domain crowd density labels to regularize target-domain crowd density distribution.

In summary, the contributions are organized as follows:

- We propose to treat crowd and background differently and design a crowd-aware mechanism for domain adaptive crowd counting.
- We propose a simple and effective schema to derive segmentation from point-level crowd counting annotations. Two crowd-aware domain adaptation modules are further proposed, based on point-derived segmentation, to guide crowd features transfer across domains beyond back-ground distraction and regularize target-domain crowd density generation.
- Our method outperforms previous approaches consistently in the widely-used adaptation scenarios.

II. RELATED WORK

Domain-adaptive Crowd Counting. Recently, some methods are proposed to solve domain-adaptive crowd counting. They can be mainly grouped into three categories. (i) Pixel-level adaptation methods [6], [9] constructs a synthetic dataset GCC and modifies CycleGAN [14] to conduct style transfer to generate target-domain crowd images for supervised training. (ii) Feature-level adaptation methods: Gao et al. [10] propose to discriminate features across domains and constrain density map generation by source-domain density labels. Han et al. [9] constrain the feature extraction by a feature discriminator and an auxiliary semantic task. Hossain et al. [8] reduce the domain shift by minimizing the feature distances (i.e., Maximum Mean Discrepancy (MMD) [15]) across domains. (iii) Others [16]–[19]: [16] introduce an extra head detector for mutual training with the crowd counter. [17] present a neuron linear transformation to optimize a small amount of for mutual training with the crowd counter. [17] present a (iii) introduce an external template encoding domain-specific meta information for humans. [19] exploit a density isomorphism reconstruction objective derived from consecutive frames in crowd videos. Methods in others can be regarded as supplements with additional bounding box annotations [16], extra target-domain annotations [17], an external template encoding [18], or temporal consistency in videos [19].

While effective, they all conduct domain adaptation in a holistic manner. However, domain alignment between crowd and background inevitably incurs misalignment, leaving room for improvement of previous methods.

Crowd Counting and Domain Adaptation. Due to limited space, generic crowd counting and domain adaptation methods are discussed in the Appendix.

III. METHOD

A. Problem Formulation

In domain adaptive crowd counting, we are given a labeled source domain \( D_S = \{ (x_i^s, y_i^s) \}_{i=1}^{N_S} \) where \( x_i^s \) and \( y_i^s \) denote the \( i \)-th crowd image and the corresponding annotation, i.e., coordinates of head positions. Besides, we have access to a unlabeled target domain \( D_T = \{ (x_i^t) \}_{i=1}^{N_T} \). Our goal is to improve counting performance in the unlabeled target domain \( D_T \) utilizing knowledge from both domains.

B. Framework Overview

As shown in Fig. 2, we propose a Crowd-aware domain Adaptation framework for Crowd Counting (CACC), which contains a crowd counter, a Point-derived Crowd Segmentation (PCS) network, and two crowd-aware adaptation modules, i.e., Crowd Region Transfer (CRT) and Crowd Density Alignment (CDA). Details of the basic crowd counter are in the Appendix.

C. Point-derived Crowd Segmentation

Point-derived Crowd Segmentation (PCS) is proposed to disentangle crowd from background by point-level crowd counting annotations in a weakly-supervised manner. The rationale behind this design is that although point annotations do not specify segmentation, they still entail where crowd appears and how crowd looks from a statistical perspective. This is also studied in the context of Multiple Instance Learning (MIL) [20] where a label is assigned to each bag of instances instead of each instance. In our case, each patch cropped from crowd images can be regarded as a bag of pixels where patch-level labels can be defined as follows.

Specifically, we densely sample patches from crowd images to construct crowd or background bags \( B = \{ b_1, b_2, ..., b_N \} \). Each patch \( b_i \) in \( B \) contains a set of pixels \( X_i = \{ x_1, x_2, ..., x_{h_i \times w_i} \} \). Let \( y_j \) be the label of each pixel \( x_j \) which indicates whether it is annotated in crowd counting. Following the standard MIL assumption that a negative bag contains only negative instances while a positive bag contains at least one positive instance, we partition \( B \) into crowd bags \( B_C \) and background bags \( B_B \) according to whether a bag contains at least a crowd counting annotation or not:

\[
B_C = \{ b_i \in B | \exists j, y_j = 1 \}, \quad B_B = \{ b_i \in B | \forall j, y_j = 0 \}.
\]  

To learn segmentation from patch-level labels, we build a learner \( \mathcal{F} \) which classifies crowd and background patches. Given each sample \( b_i \) from \( B_C \) or \( B_B \), \( \mathcal{F} \) outputs an intermediate 2-channel map \( M_i = \mathcal{F}(b_i, \Theta) \). Optimization objective of classifier \( \mathcal{F} \) is a standard cross entropy loss:

\[
L_F = \sum_{b_i \in B_C} -\log(\mathcal{S}(\mathcal{A}(M_i))) + \sum_{b_i \in B_B} -\log(\mathcal{S}(\mathcal{A}(M_i))),
\]  

where \( \mathcal{A}() \) is a 2D aggregator (e.g., Avg2D), \( \mathcal{S}(\cdot) \) is the softmax function. \( M_i \) and \( M_i \) represent the first and second channel of \( M_i \). The learning of \( \mathcal{F} \) can activate pixel-wise
Fig. 2: Overview of the proposed Crowd-aware domain Adaptation framework for Crowd Counting (CACC). To disentangle crowd from background, we derive crowd segmentation from point-level crowd counting annotations, namely Point-derived Crowd Segmentation (PCS), in a weakly-supervised manner. Based on the derived segmentation, we propose two crowd-aware adaptation modules, i.e., Crowd Region Transfer (CRT) and Crowd Density Alignment (CDA). Crowd Region Transfer guides crowd features alignment across domains beyond background distractions. Crowd Density Alignment samples pseudo head positions from segmentation to generate segmentation-guided pseudo labels, which are utilized to regularize target-domain crowd density generation by its own crowd density distribution.

D. Crowd Region Transfer

Crowd Region Transfer (CRT) is designed to align crowd features across domains beyond background distractions by learning domain-invariant crowd feature representations.

Given a crowd image x, we denote crowd segmentation from Point-derived Crowd Segmentation as C_{seg}. We have two variants to design segmentation, i.e., soft crowd segmentation C_{seg}^S and hard crowd segmentation C_{seg}^H. We directly utilize C_{seg}^S as C_{seg}. For C_{seg}^H, we binarize C_{seg} by:

\[ T = \frac{1}{HW} \sum_{h,w} C_{seg}^S(h,w) \cdot C_{seg}^H = \mathcal{I}(C_{seg}^S > T), \]  

where threshold T is set to the mean value of C_{seg}^S, \( \mathcal{I}(\cdot) \) represents the indication function.

Following [9]–[11], we utilize adversarial training to learn domain-invariant features. Differently, we discard domain-specific background features and focus on domain-invariant crowd features. Optimization objective of CRT is:

\[
\mathcal{L}_{\text{CRT}} = \min_{\theta_G} \max_{\theta_D} \mathbb{E}_{x_s \sim \mathcal{D}_S} \log D(G(x_s) \cdot C_{seg}^H) + \mathbb{E}_{x_t \sim \mathcal{D}_T} \log (1 - D(G(x_t) \cdot C_{seg}^H)),
\]

where G is the feature extractor of crowd counter, D is domain classifier. D and G construct a two-player minimax game, where D is trained to distinguish which domain the features come from, while G aims to confuse D.

Note that soft crowd segmentation C_{seg}^S from PCS is head-highlighted. When utilized in Eq. 4, C_{seg}^S enhances head features alignment across domains, which is crucial for crowd counting due to its head-centric recognition mechanism. Effectiveness of C_{seg}^S is shown in Table I.
Algorithm 1: Crowd-aware Domain Adaptation for Crowd Counting.

Input: Labeled source domain $D_S$, Unlabeled target domain $D_T$, Batch size $B$.

Output: A domain adaptive crowd counter $C(\cdot, \theta)$.

1. Supervised learning of $C(\cdot, \theta)$ in $D_S$.
2. Sample bags $B$ in $D_S$ & $D_T$ and partition $B$ into $B_C$ and $B_B$ by Eq. (1).
3. Learning of $F$ on $B_C$ and $B_B$ by Eq. (2).
4. Obtain crowd segmentation in $D_S$ and $D_T$.
5. for $i = 1$ to max_iter do
   6. $X_S, Y_S \leftarrow \text{Sample}(D_S, B/2)$
   7. $X_T \leftarrow \text{Sample}(D_T, B/2)$
   8. Calculate $L_{den}$
   9. Calculate $L_{CRT}$ by Eq. (4).
   10. Generate segmentation-guided pseudo labels in $D_T$
   11. Calculate $L_{CDA}$ according to Eq. (6).
   12. Optimize $C(\cdot, \theta)$ by Eq. (7).

E. Crowd Density Alignment

Crowd Density Alignment (CDA) is designed to regularize target-domain crowd density generation by its own crowd density distribution, instead of source-domain density distribution utilized in all the previous methods.

Specifically, we use soft crowd segmentation $C_{seg}^S$ to generate probabilistic crowd distribution $P$ by normalization:

$$P = \frac{C_{seg}^S}{\sum_{h,w} C_{seg}^S(h,w)}.$$  (5)

$P$ follows a discrete bivariate distribution where we iteratively sample pseudo head positions $P = \{(w_i, h_i) | i \in [1, n]\}$. After sampling, we generate pseudo density labels as in the source domain by convolving each pseudo head point with a Gaussian kernel.

Following previous methods, we utilize adversarial training to regularize target-domain crowd density generation. Differently, we exploit segmentation-guided pseudo density labels as guidance, instead of source-domain density labels. The optimization objective is:

$$L_{CDA} = \min_{\theta_C} \max_{\theta_D} \mathbb{E}_{M_{SPL} \sim D_{SPL}} \log D_m(M_{SPL})$$
$$+ \mathbb{E}_{X_t \sim D_T} \log(1 - D_m(G(x_t))),$$  (6)

where $D_m$ denotes crowd density discriminator. $M_{SPL}$ and $D_{SPL}$ represent segmentation-guided pseudo density maps and the corresponding domain respectively. With the segmentation-guided pseudo labels, our method can directly constrain the target-domain crowd density generation by its own crowd density distribution.

Note that the head-highlighted nature of soft crowd segmentation $C_{seg}^S$ also benefits the generation of segmentation-guided pseudo labels considering the head-centric labeling mechanism of crowd counting.

F. Network Optimization

The training procedure of the proposed framework contains three major components: Supervised Learning (SL) $L_{den}$, Crowd Region Transfer (CRT) $L_{CRT}$, and Crowd Density Alignment (CDA) $L_{CDA}$. With the above terms, the overall optimization objective writes as:

$$L_{total} = L_{den} + \lambda_1 L_{CRT} + \lambda_2 L_{CDA},$$  (7)

where $\lambda_1$ and $\lambda_2$ are factors to balance the three items. Detailed optimization procedure is shown in Algorithm 1.

IV. EXPERIMENTS

A. Datasets and Adaptation Scenarios

Datasets. Six datasets are used in our experiments, i.e., GCC [6], ShanghaiTech PartA (SHPartA) [4], ShanghaiTech PartB (SHPartB) [4], JHU-CROWD (JHUC) [21], MALL [22], and UCSD [23]. Details are in the Appendix.

Adaptation Scenarios. (i) Synthetic-to-Real (GCC → SHPartB, GCC → SHPartA). We employ the synthetic GCC as source domain and the training set of SHPartB or SHPartA as target domain. (ii) Fixed-to-Fickle (SHPartB → SHPartA). We utilize the training set of SHPartB (a fixed crowd scene) as source domain and the training set of SHPartA (various crowd scenes) as target domain. (iii) Normal-to-BadWeather (SHPartA → JHUC). To simulate weather condition changes, we utilize the training set of SHPartA as source domain and the images with bad weather conditions in the training set of JHUC as target domain.

B. Ablation Studies

We conduct ablation studies in Synthetic-to-Real adaptation scenario to validate the effectiveness of the proposed modules, i.e., PCS, CRT, and CDA.

| Method                | GCC → SHPartB | GCC → SHPartA |
|-----------------------|---------------|---------------|
| Source only           | 19.5          | 169.2         |
| CRT w/o PCS           | 16.4          | 154.9         |
| CRT w/ BinarySeg.     | 15.6          | 123.8         |
| CRT w/ Hard Seg.      | 15.0          | 204.9         |
| CRT w/ Soft Seg.      | 14.7          | 203.2         |
| CRT w/o PCS           | 26.8          | 255.9         |
| CRT w/ BinarySeg.     | 24.1          | 213.6         |
| CRT w/ Hard Seg.      | 23.8          | 204.9         |
| CRT w/ Soft Seg.      | 23.5          | 201.6         |

Table I: Ablation studies on Crowd Region Transfer (CRT) in the Synthetic-to-Real adaptation scenario.
Effectiveness of PCS. We evaluate PCS by testing how much the derived crowd segmentation covers annotated human heads. The percentages of coverage in Synthetic-to-Real adaptation scenario are 98.5, 93.6, and 95.2 for GCC (source domain), SHPartA (target domain), and SHPartB (target domain) datasets, respectively. This indicates that point-derived crowd segmentation can cover almost all human heads in both domains. Qualitative results of PCS are shown in Fig. 4.

Effectiveness of CRT. To evaluate the effectiveness of CRT, we introduce several comparison variants as follows. “Source only” denotes crowd counter trained on source domain only. “CRT w/o PCS” transfers features across domains in a holistic manner. “CRT w/ Hard Seg.”, “CRT w/ Soft Seg.”, and “CRT w/ BinarySeg.” denote CRT with hard crowd segmentation, soft crowd segmentation, and binarizing Gaussian-blurred density maps [24].

As shown in Table I compared to “Source Only”, “CRT w/o PCS” can improve the adaptation performance to some extent. “CRT w/ BinarySeg.”, “CRT w/ Hard Seg.”, and “CRT w/ Soft Seg.” achieve lower counting errors compared to “CRT w/o PCS” no matter what kind of crowd segmentation is leveraged. This indicates background features alignment across domains incurs an adverse effect during domain adaptation. “CRT w/ Soft Seg.” is better than “CRT w/ Hard Seg.”, which demonstrates the effectiveness of enhanced head features brought by head-highlighted soft crowd segmentation.

Effectiveness of CDA. As shown in Table VII “CDA” outperforms “SL” (Source-domain density Labels) [10] consistently, which demonstrates the superiority of the segmentation-guided density alignment mechanism. Qualitative results of segmentation-guided pseudo labels are in Fig. 4.

C. Comparison to state-of-the-art methods

Synthetic-to-Real. As shown in Table II our method can achieve the lowest counting errors and the highest relative gains compared to all the unsupervised counterparts. Although we do not leverage extra annotations, our method can still outperform FSC [9]. To be comparable with NLT [17], we also introduce 10% labeled target data. The performance of our method in terms of MAE/RMSE is enhanced to 10.2/17.5 (SHPartB) and 82.4/136.6 (SHPartA), respectively, which are also better than NLT [17].

Fixed-to-Fickle & Normal-to-BadWeather. The two adaptation scenarios are discussed for the first time in the literature. Due to limited space, we show the results in the Appendix.

Others. To conduct more comparisons with state-of-the-art methods, we follow some other shared settings, i.e., SHPartA→SHPartB, MALL→UCSD, and UCSD→MALL. As can be seen in Table IV our method can outperform state-of-the-art methods in different adaptation scenarios.

D. Qualitative Results

Qualitative results of the estimated density maps can be seen in Fig. 5. Due to the domain shift problem, the “Source Only” model simply detects some salient individuals in the crowd. From “Source Only” to “Ours w/o PCS”, we can observe that the “Ours w/o PCS” model can increase true positives to some extent, but also incurs some false positives in the background areas due to the misalignment between crowd and background. Differently, our method can consistently estimates more accurate crowd densities and suppresses the occurrence of false positives thanks to the proposed crowd-aware domain adaptation method.
V. CONCLUSION

In this paper, we propose to treat crowd and background differently and design a Crowd-aware domain Adaptation framework for Crowd Counting (CACC). Specifically, we learn crowd segmentation from pixel-level crowd counting annotations. Based on the derived segments, we design two crowd-aware adaptation modules, i.e., Crowd Region Transfer (CRT) and Crowd Density Alignment (CDA). Extensive experiments in multiple cross-domain scenarios demonstrate the superiority of the proposed method.

REFERENCES

[1] Teng Li, Huan Chang, Meng Wang, Bingbing Ni, Richang Hong, and Shuicheng Yan, “Crowded scene analysis: A survey,” IEEE transactions on circuits and systems for video technology, vol. 25, no. 3, pp. 367–386, 2014.

[2] Guangshuai Gao, Junyu Gao, Qingjie Liu, Qi Wang, and Yunhong Wang, “Cnn-based density estimation and crowd counting: A survey,” arXiv preprint arXiv:2003.12783, 2020.

[3] Antonio Torralba and Alexei A Efros, “Unbiased look at dataset bias,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[4] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma, “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.

[5] Sinno Jialin Pan and Qiang Yang, “A survey on transfer learning,” IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2009.

[6] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan, “Learning from synthetic data for crowd counting in the wild,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[7] Junyu Gao, Tao Han, Yuan Yuan, and Qi Wang, “Domain-adaptive crowd counting via high-quality image translation and density reconstruction,” IEEE Transactions on Neural Networks and Learning Systems, 2021.

[8] Mohammad Asifal Hosain, Mahesh Kumar Krishna Reddy, Kevin Cannons, Zhan Xu, and Yang Wang, “Domain adaptation in crowd counting,” in IEEE Conference on Computer and Robot Vision, 2020.

[9] Tao Han, Junyu Gao, Yuan Yuan, and Qi Wang, “Focus on semantic consistency for cross-domain crowd understanding,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2020.

[10] Junyu Gao, Qi Wang, et al., “Feature-aware adaptation and density alignment for crowd counting in video surveillance,” IEEE Transactions on Cybernetics, 2020.

[11] Zhikang Zou, Xiaoye Qu, Pan Zhou, Shuangjie Xu, Xiaojing Ye, Wenhao Wu, and Jin Ye, “Coarse to fine: Domain adaptive crowd counting via adversarial scoring network,” in Proceedings of the ACM International Conference on Multimedia, 2021, pp. 2185–2194.

[12] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei, “Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation,” in European Conference on Computer Vision, 2020.

[13] Sinan Wang, Xinyang Chen, Yunbo Wang, Mingsheng Long, and Jianmin Wang, “Progressive adversarial networks for fine-grained domain adaptation,” in CVPR, 2020.

[14] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proceedings of the IEEE International Conference on Computer Vision, 2017.

[15] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan, “Deep transfer learning with joint adaptation networks,” in International Conference on Machine Learning, 2017.

[16] Yuting Liu, Zheng Wang, Miaoqin Shi, Shin’ichi Satoh, Qijun Zhao, and Hongyu Yang, “Towards unsupervised crowd counting via regression-detection bi-knowledge transfer,” in Proceedings of the 28th ACM International Conference on Multimedia, 2020.

[17] Qi Wang, Tao Han, Junyu Gao, and Yuan Yuan, “Neuron linear transformation: modeling the domain shift for crowd counting,” IEEE Transactions on Neural Networks and Learning Systems, 2021.

[18] Qi Wang, Tao Han, Junyu Gao, and Yuan Yuan, “Neutron linear transformation: modeling the domain shift for crowd counting,” in Proceedings of the ACM International Conference on Multimedia, 2021, pp. 658–666.

[19] Yuhang He, Zhiheng Ma, Xing Wei, Xiaopeng Hong, Wei Ke, and Yihong Gong, “Error-aware density isomorphism reconstruction for unsupervised cross-domain crowd counting,” in AAAI, 2021.

[20] Oded Maron and Tomás Lozano-Pérez, “A framework for multiple-instance learning,” Advances in neural information processing systems, pp. 570–576, 1998.

[21] Vishwanath Sindagi, Rajeev Yasarla, and Vishal MM Patel, “Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method,” IEEE transactions on pattern analysis and machine intelligence, 2020.

[22] Ke Chen, Chen Chang Eloy, Shaogang Gong, and Tony Xiang, “Feature mining for localised crowd counting,” in BMVC, 2012, vol. 1, p. 3.

[23] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos, “Privacy preserving crowd monitoring: Counting people without people models or tracking,” in 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–7.

[24] Yan Liu, Lingqiao Liu, Peng Wang, Pingping Zhang, and Yinjie Lei, “Semi-supervised crowd counting via self-training on surrogate tasks,” in European Conference on Computer Vision, 2020.

[25] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah, “Multi-source multi-scale counting in extremely dense crowd images,”
Yaroslav Ganin and Victor Lempitsky, “Unsupervised domain adaptation by backpropagation,” in International conference on machine learning, 2015.

Yaroslav Ganin, Evguenia Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, “Domain-adversarial training of neural networks,” The Journal of Machine Learning Research, vol. 17, no. 1, pp. 2096–2030, 2016.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, “Adversarial discriminative domain adaptation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

Qingchao Chen, Yang Liu, Zhaowen Wang, Ian Wassell, and Kevin Chetty, “Re-weighted adversarial adaptation network for unsupervised domain adaptation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu, “Collaborative and adversarial network for unsupervised domain adaptation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

Riccardo Volpi, Pietro Morerio, Silvio Savarese, and Vittorio Murino, “Adversarial feature augmentation for unsupervised domain adaptation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang, “Progressive feature alignment for unsupervised domain adaptation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019.

Yabin Zhang, Hui Tang, Kii Jia, and Mingkui Tan, “Domain-symmetric networks for adversarial domain adaptation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019.

Seungmin Lee, Dongwan Kim, Namil Kim, and Seong-Gyun Jeong, “Drop to adapt: Learning discriminative features for unsupervised domain adaptation,” in Proceedings of the IEEE international conference on computer vision, 2019.

Minghao Chen, Shaui Zhao, HaiFeng Liu, and Deng Cai, “Adversarial-learned loss for domain adaptation,” in Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenzhen Zhang, “Adversarial domain adaptation with domain mixup,” in Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada, “Asymmetric tri-training for unsupervised domain adaptation,” in International conference on machine learning, 2017.

Zhengming Ding, Sheng Li, Ming Shao, and Yun Fu, “Graph adaptive knowledge transfer for unsupervised domain adaptation,” in European Conference on Computer Vision, 2018.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann, “Contrastive adaptation network for unsupervised domain adaptation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019.

Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa, “Learning from synthetic data: Addressing domain shift for semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Mannohandre Chandraker, “Learning to adapt structured output space for semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

Yawei Zhang, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang, “Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019.

Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiayang Yang, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang, “Sifsdan: Separated semantic feature based domain adaptation network for semantic segmentation,” in Proceedings of the IEEE international conference on computer vision, 2019.

Jihan Yang, Ruijia Xu, Ruyi Li, Xiaojuan Qi, Xiaoyong Shen, Guanbin Li, and Liang Lin, “An adversarial perturbation oriented domain adaptation approach for semantic segmentation,” in Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Wen-mei Hwu, Thomas S Huang, and Honghui Shi, “Different treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2020.

Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon, “Unsupervised intra-domain adaptation for semantic segmentation through self-supervision,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool, “Domain adaptive faster r-cnn for object detection in the wild,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changikk Kim, “Diversify and match: A domain adaptive representation learning paradigm for object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019.

Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko, “Strong-weak distribution alignment for adaptive object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019.

Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin, “Adapting object detectors via selective cross-domain alignment,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019.

Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao, “Exploring object relation in mean teacher for cross-domain detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019.

Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang, “Cross-domain object detection through coarse-to-fine feature adaptation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2020.

Vishwath A Sindagi, Poojan Oza, Rajeev Yasarla, and Vishal M Patel, “Prior-based domain adaptive object detection for hazy and rainy conditions,” in European Conference on Computer Vision, 2020.

Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

VI. APPENDIX

A. More Related Work

1) Crowd Counting: Early works for crowd counting are mainly based on hand-crafted features (e.g., SIFT, Fourier Analysis, HOG) to estimate crowd counts by either regression [22], [25], [26] or object detection [27]–[29]. Various CNN-based methods have advanced the performance of crowd counting. Most of them are dedicated to handle various challenges of crowd counting in an supervised manner, e.g., large scale variations [4], [30]–[37], hand-crafted gaussian kernels [38]–[40], uncertainty [41], [42], enhancing crowd features [43]–[46], extra constraints [47]–[49], etc. Besides the supervised methods, several approaches focus on relieving the labeling burdensome. They can be broadly categorized into semi-supervised methods [24], [50]–[54], weakly-supervised methods [55], [56], self-supervised methods [57], [58] and unsupervised methods [59], [60].

These generic crowd counters can achieve promising performance in public datasets, whereas they do not focus on solving the domain shift problem, which hurts their generalization performance in real-world application scenarios.

2) Domain Adaptation: Lots of domain adaptation methods dedicate to reducing domain discrepancies by learning domain-invariant feature representations. Methods along this
TABLE V: Architecture of crowd counter.

| VGG16 Backbone |
|----------------|
| Conv1: [K(3,3)-C64-S1-R] |
| ... |
| Conv10: [K(3,3)-C512-S1-R] |

Deconvolution Block

| Conv11: [K(3,3)-C64-S1-R]; Deconv1: [K(2,2)-C64-S2-R] |
| Conv12: [K(3,3)-C32-S1-R]; Deconv2: [K(2,2)-C32-S2-R] |
| Conv13: [K(3,3)-C16-S1-R]; Deconv3: [K(2,2)-C16-S2-R] |

Density Regression Layer

| Conv14: [K(3,3)-C16-S1-R] |
| Conv15: [K(3,3)-C1-S1-R] |

TABLE VI: Architecture of Point-derived Crowd Segmentation (PCS) network.

| Feature Extractor |
|-------------------|
| Conv1: [K(3,3)-C64-S1-R]; Conv2: [K(3,3)-C64-S1-R] |
| MaxPool1: [K(2,2)-C16-S2] |
| Conv3: [K(3,3)-C32-S1-R]; Conv4: [K(3,3)-C32-S1-R] |
| MaxPool2: [K(2,2)-C32-S2] |
| Conv5: [K(3,3)-C32-S1-R]; Conv6: [K(3,3)-C32-S1-R] |
| MaxPool3: [K(2,2)-C32-S2] |
| Conv7: [K(3,3)-C2-S1-R] |

2DAvgPool & Softmax

line can be generally categorized into two types: criterion-based methods [15], [61]–[66] and adversarial learning-based methods [47]–[49]. The former aligns feature distributions between different domains by minimizing some statistics, such as Maximum Mean Discrepancy [15], Correlation Alignment [62], Wasserstein distance [65], and HoMM [66]. The latter introduces a domain discriminator to classify feature representations, while adversarially confuses the discriminator by constructing a minimax game with the feature extractor. These methods have been widely studied and achieved superior performance in image classification [67], [78]–[80], semantic segmentation [81], [82], object detection [83], [84].

However, domain adaptation for crowd counting is less studied, and existing generic methods cannot easily adapt to crowd counting due to its special labeling mechanism and diverse backgrounds in crowd scenes.

B. More Network Details

1) Architecture of Crowd Counter: Most crowd counting networks employ density maps as the intermediate output for better supervision. They are typically generated by convolving each annotated head point with a Gaussian kernel [4]:

\[
D(z) = \sum_{k=1}^{N} \delta(z - z_k) * G_{\sigma_k}(z), \quad (8)
\]

where \( z \) and \( z_k \) denote each pixel and the \( k \)-th annotated point (total \( N \) points) in a crowd image \( x \). \( G_{\sigma_k} \) is a 2D Gaussian kernel with a bandwidth \( \sigma_k \). Following previous works [6], [9], [10], we employ a simple and universal crowd counter without specialized techniques to verify the general effectiveness of the proposed domain adaptation method. Specifically, we extract the first ten convolutional layers of VGG16 [5] with three maxpooling layers as the backbone network. After the backbone network, we introduce several deconvolutional layers to generate high-resolution density maps. Detailed network architecture of crowd counter is in Table V. For example, “K(3,3)-C64-S1-R” represents the Convolution or Deconvolution layer and \( \sigma \) is a 2D Gaussian kernel with a bandwidth \( \sigma_k \) for multi-scale features extracted after each pooling layer in the backbone network.

CDA for multi-scale features extracted after each pooling layer in the backbone network.

2) Datasets: Six datasets are utilized in our experiments. (i) GCC [6] is a synthetic dataset containing 15,212 images with resolution of 1080 × 1920, which are rendered by GTA5 and captured by 400 surveillance cameras in a fictional city. (ii) SHPartA [4] is randomly crawled from the Internet with various crowd scenes containing 482 images, in which 300 images for training and 182 images for testing. (iii) SHPartB [4] is collected from the busy streets of metropolitan areas in Shanghai consisting of 716 images, in which 300 images for training and the remaining for testing. Compared to SHPartA, SHPartB has relatively fixed camera perspectives.
and crowd scenes. (iv) **JHU-CROWD (JHUC)** [21] is a large-scale dataset proposed recently, which contains 4,372 images. Images are collected under a variety of scenes and environmental conditions, and annotations include head positions, approximate sizes, blur-level, occlusion-level, weather-labels, etc. (v) **MALL** [22] is captured in a shopping mall by a fixed surveillance camera. The dataset consists of 2,000 frames in which the first 800 frames for training and the remaining for testing. (vi) **UCSD** [23] is collected by a fixed video camera besides a pedestrian walkway. The datasets contains 2,000 frames in which the training set captures 601 to 1,400 and the testing set owns the remaining. Region-of-interest (ROI) and perspective map are provided.

3) **More Quantitative Results:** **Effectiveness of CDA.** To further verify the effectiveness of Crowd Density Alignment (CDA), we conduct more ablation studies without based on the proposed Crowd Region Transfer (CFA). The comparison results are in Table VII. We can see that “CDA” outperforms “SL”, which demonstrates consistent superiority of the proposed segmentation-guided density alignment mechanism.

**Fixed-to-Fickle & Normal-to-BadWeather.** The two adaptation scenarios are discussed for the first time in the literature. However, they are also very important adaptation scenarios considering various crowd scenes and weather conditions in real-world applications. Results of different variants of our method in the two scenarios are summarized in Table VIII. As can be seen, the proposed PCS, CRT, and CDA modules can progressively improve the counting accuracies in both adaptation scenarios, which confirms the effectiveness of the proposed crowd-aware domain adaptation mechanism in multiple real-world adaptation scenarios.