7-2023

Fine-grained domain adaptive crowd counting via point-derived segmentation

Yongtuo LIU

Dan XU

Sucheng REN

Hanjie WU

Hongmin CAI

See next page for additional authors

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

Yongtuo Liu[1,2], Dan Xu[3], Sucheng Ren[1], Hanjie Wu[1], Hongmin Cai[1], Shengfeng He[1,4*]

[1]School of Computer Science, South China University of Technology
[2]Institute of Informatics, University of Amsterdam
[3]Department of Computer Science and Engineering, Hong Kong University of Science and Technology
[4]School 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.**

***Index Terms*—Crowd Counting, Domain Adaptation, Point-derived Segmentation**

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.

*Shengfeng He is the corresponding author (shengfenghe7@gmail.com)

(a) Labeled Source Domain *SHPartB* **(b) Unlabeled Target Domain *SHPartA***

Fig. 1: Crowd images from *SHPartB* (a) and *SHPartA* (b) datasets [4] respectively. As can be seen, backgrounds vary a lot across domains. For instance, backgrounds in *SHPartB* are mainly ground whereas various backgrounds appear in *SHPartA* including buildings, trees, sky, etc.

While achieving promising results, existing domain adaptative 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 background 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]: [6] 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 parameters based on a few target-domain training samples. [18] 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 $\mathcal{D}_{\mathcal{S}} = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{N_s}$ where $\mathbf{x}_i^s$ and $\mathbf{y}_i^s$ 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 $\mathcal{D}_{\mathcal{T}} = \{(\mathbf{x}_i^t)\}_{i=1}^{N_t}$. Our goal is to improve counting performance in the unlabeled target domain $\mathcal{D}_{\mathcal{T}}$ 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 $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_N\}$. Each patch $\mathbf{b}_i$ in $\mathcal{B}$ contains a set of pixels $\mathbf{X}_i = \{x_1, x_2, ..., x_{h_i \times w_i}\}$. 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 $\mathcal{B}$ into crowd bags $\mathcal{B}_C$ and background bags $\mathcal{B}_B$ according to whether a bag contains at least a crowd counting annotation or not:

$$\begin{aligned} \mathcal{B}_C &= \{\mathbf{b}_i \in \mathcal{B} \; if \; y_j = 1, \exists \; x_j \in \mathbf{b}_i\}, \\ \mathcal{B}_B &= \{\mathbf{b}_i \in \mathcal{B} \; if \; y_j = 0, \forall \; x_j \in \mathbf{b}_i\}. \end{aligned} \tag{1}$$

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

$$\begin{aligned} \mathcal{L}_{\mathcal{F}} = &\sum_{\mathbf{b}_i \in \mathcal{B}_C} -\log(\mathcal{S}(\mathcal{A}(\mathbf{M}_i^0))) \\ &+ \sum_{\mathbf{b}_i \in \mathcal{B}_B} -\log(\mathcal{S}(\mathcal{A}(\mathbf{M}_i^1))), \end{aligned} \tag{2}$$

where $\mathcal{A}(\cdot)$ is a 2D aggregator (e.g., Avg2D), $\mathcal{S}(\cdot)$ is the softmax function. $\mathbf{M}_i^0$ and $\mathbf{M}_i^1$ represent the first and second channel of $\mathbf{M}_i$. The learning of $\mathcal{F}$ can activate pixel-wise

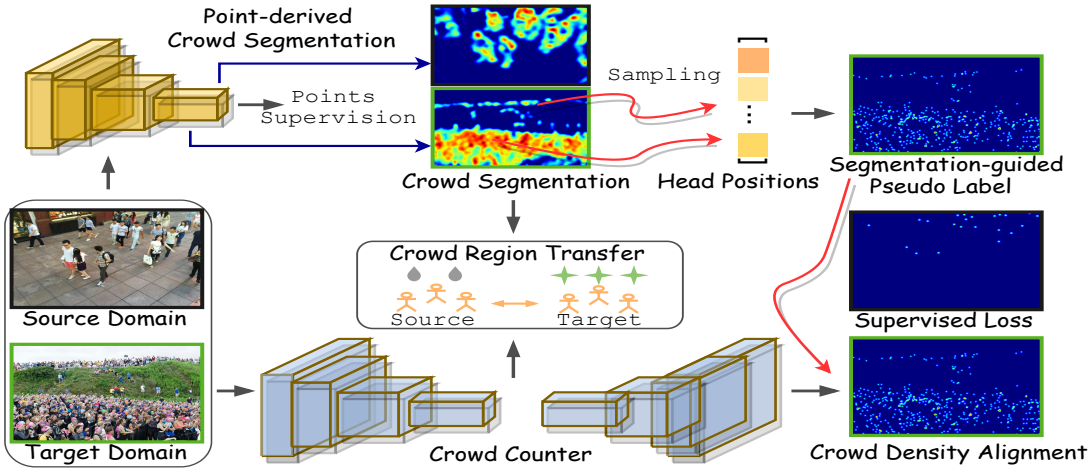

Fig. 2: Overview of the proposed Crowd-aware domain Adaptation framework for Crowd Counting (CACC). To disentangle crowd from background, we derive crowd segmentation from point-level crowd counting annotations, namely Point-derived Crowd Segmentation (PCS), in a weakly-supervised manner. Based on the 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.

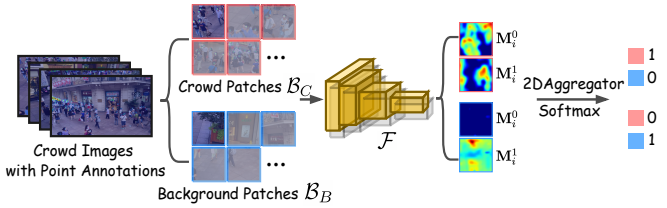

Fig. 3: Learning Point-derived Crowd Segmentation (PCS) from point-level crowd counting annotations. 0 and 1 in each side constitute a one-hot vector indicating the patch is annotated as a crowd or background patch.

responses in $\mathbf{M}_i$ for better discrimination of crowd and background patches. After learning converges, we utilize $\mathbf{M}_i$ as crowd and background segmentation. PCS is shown in Fig. 3.

Note that crowd counting annotates human head positions only. A background bag $\mathcal{B}_B$ in Eq. (1) may contain human bodies. However, $\mathcal{F}$ in Eq. (2) is trained statistically with natural tolerance to noisy labels. As shown in Fig. 2, we can still derive high-quality crowd segmentation from noisy labels. As human heads only exist in crowd bags $\mathcal{B}_C$, activated crowd segmentation can thus focus more on human heads for better classification. This is a blessing in disguise when head-highlighted crowd segmentation is utilized to design crowd-aware domain adaptation modules. This is because of the head-centric labeling and recognition nature of crowd counting. We will detail the benefits of head-highlighted crowd segmentation in the following proposed modules.

In practice, only the source domain has crowd counting annotations. We utilize density maps estimated by crowd counter for training of $\mathcal{F}$ in the target domain. As discussed above, $\mathcal{F}$ does not rely on accurate labels due to its statistical nature. As shown in Fig. 4, segmentation results in the target domain are still ensured high quality.

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 $\mathbf{x}$, we denote crowd segmentation from Point-derived Crowd Segmentation as $\mathbf{C}_{seg}$. We have two variants to design segmentation, i.e., soft crowd segmentation $\mathbf{C}^S_{seg}$ and hard crowd segmentation $\mathbf{C}^H_{seg}$. We directly utilize $\mathbf{C}_{seg}$ as $\mathbf{C}^S_{seg}$. For $\mathbf{C}^H_{seg}$, we binarize $\mathbf{C}^S_{seg}$ by:

$$T = \frac{1}{HW}\sum_{h,w}\mathbf{C}^S_{seg}(h,w),\ \ \mathbf{C}^H_{seg} = \mathcal{I}(\mathbf{C}^S_{seg} > T), \tag{3}$$

where threshold T is set to the mean value of $\mathbf{C}^S_{seg}$, $\mathcal{I}(\cdot)$ 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:

$$\begin{aligned}\mathcal{L}_{CRT} = \min_{\theta_G}\max_{\theta_D} &\mathbb{E}_{\mathbf{x}_s\sim\mathcal{D}_\mathcal{S}}\log D(G(\mathbf{x}_s)\cdot\mathbf{C}^{H/S}_{seg}) \\ &+ \mathbb{E}_{\mathbf{x}_t\sim\mathcal{D}_\mathcal{T}}\log(1 - D(G(\mathbf{x}_t)\cdot\mathbf{C}^{H/S}_{seg})),\end{aligned} \tag{4}$$

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 $\mathbf{C}^S_{seg}$ from PCS is head-highlighted. When utilized in Eq. 4, $\mathbf{C}^S_{seg}$ enhances head features alignment across domains, which is crucial for crowd counting due to its head-centric recognition mechanism. Effectiveness of $\mathbf{C}^S_{seg}$ is shown in Table I.

Algorithm 1: Crowd-aware Domain Adaptation for Crowd Counting.

Input: Labeled source domain $\mathcal{D}_{\mathcal{S}}$.
Unlabeled target domain $\mathcal{D}_{\mathcal{T}}$.
Batch size B.
Output: A domain adaptive crowd counter $C(\cdot, \theta)$.

Supervised learning of $C(\cdot, \theta)$ in $\mathcal{D}_{\mathcal{S}}$.
Sample bags $\mathcal{B}$ in $\mathcal{D}_{\mathcal{S}}$ & $\mathcal{D}_{\mathcal{T}}$ and partition $\mathcal{B}$ into $\mathcal{B}_C$ and $\mathcal{B}_B$ by Eq. (1).
Learning of $\mathcal{F}$ on $\mathcal{B}_C$ and $\mathcal{B}_B$ by Eq. (2).
Obtain crowd segmentation in $\mathcal{D}_{\mathcal{S}}$ and $\mathcal{D}_{\mathcal{T}}$.
for $i = 1$ **to** max_iter **do**
  $X_S, Y_S \leftarrow Sample(\mathcal{D}_{\mathcal{S}}, B/2)$
  $X_T \leftarrow Sample(\mathcal{D}_{\mathcal{T}}, B/2)$
  Calculate $\mathcal{L}_{den}$
  Calculate $\mathcal{L}_{CRT}$ by Eq. (4)
  Generate segmentation-guided pseudo labels in $\mathcal{D}_{\mathcal{T}}$
  Calculate $\mathcal{L}_{CDA}$ according to Eq. (6)
  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 $\mathbf{C}^S_{seg}$ to generate probabilistic crowd distribution P by normalization:

$$P = \mathbf{C}^S_{seg} / \sum_{h,w} \mathbf{C}^S_{seg}(h, w). \tag{5}$$

P follows a discrete bivariate distribution where we iteratively sample pseudo head positions $\mathcal{P} = \{(w_i, h_i) \,|\, i \in [1, n]\}$. 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:

$$\begin{aligned}\mathcal{L}_{CDA} = \min_{\theta_G} \max_{\theta_{D_m}} & \mathbb{E}_{\mathbf{M}_{SPL} \sim \mathcal{D}_{SPL}} \log D_m(\mathbf{M}_{SPL})) \\ & + \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}_{\mathcal{T}}} \log(1 - D_m(G(\mathbf{x}_t))),\end{aligned} \tag{6}$$

where D_m denotes crowd density discriminator. $\mathbf{M}_{SPL}$ and $\mathcal{D}_{SPL}$ 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 $\mathbf{C}^S_{seg}$ also benefits the generation of segmentation-guided pseudo labels considering the head-centric labeling mechanism of crowd counting.

TABLE I: Ablation studies on Crowd Region Transfer (CRT) in the Synthetic-to-Real adaptation scenario.

Method	GCC → SHPartB		GCC → SHPartA	
	MAE	RMSE	MAE	RMSE
Source only	19.5	28.9	169.2	255.9
CRT w/o PCS	16.4	26.8	134.8	213.6
CRT w/ BinarySeg.	15.6	24.1	125.3	204.9
CRT w/ Hard Seg.	15.0	23.8	122.5	203.2
CRT w/ Soft Seg.	**14.7**	**23.5**	**117.4**	**201.6**

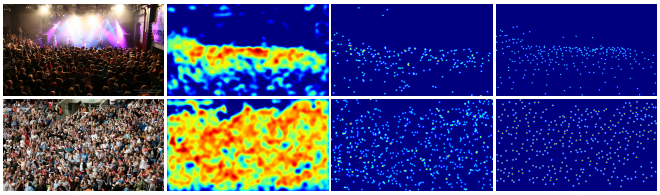

Fig. 4: Qualitative results of Point-derived Crowd Segmentation (PCS Seg.) and segmentation-guided pseudo label for Crowd Density Alignment (CDA label) in the target domains of Synthetic-to-Real adaptation scenario.

F. Network Optimization

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

$$\mathcal{L}_{total} = \mathcal{L}_{den} + \lambda_1 \mathcal{L}_{CRT} + \lambda_2 \mathcal{L}_{CDA}, \tag{7}$$

where λ_1 and λ_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.

TABLE II: Comparison with state-of-the-art methods in the Synthetic-to-Real adaptation scenario. "U" and "S" denote unsupervised and semi-supervised domain adaptation methods, respectively. "Gain" denotes the relative gains of MSE/RMSE in comparison to the performance before adaptation.

Synthetic → Real							
Method	Type	GCC → SHPartB			GCC → SHPartA		
		MAE ↓	RMSE ↓	Gain ↑	MAE ↓	RMSE ↓	Gains ↑
NLT [17]	S	10.8	18.3	46.2%/37.3%	90.1	151.6	52.0%/45.7%
FSC [9]	S	16.9	24.7	31.1%/26.7%	129.3	187.6	32.2%/37.0%
CycleGAN [14]	U	25.4	39.7	-10.2%/-23.6%	143.3	204.3	10.4%/5.6%
SE CycleGAN [6]	U	19.9	28.3	12.7%/7.5%	123.4	193.4	22.8%/10.6%
FADA [10]	U	16.0	24.7	28.2%/17.3%	–	–	–
ASNet [11]	U	14.6	22.6	–	–	–	–
Ours	U	**13.5**	**21.8**	**30.7%/24.5%**	**109.3**	**187.1**	**35.4%/26.8%**
Oracle	–	8.9	15.3	–	67.5	112.1	–

TABLE III: Ablation studies on Crowd Density Alignment (CDA) in the Synthetic-to-Real adaptation scenario. CRT here utilizes soft crowd segmentation ("CRT w/ Soft Seg.").

Method	GCC → SHPartB		GCC → SHPartA	
	MAE	RMSE	MAE	RMSE
CRT	14.7	23.5	117.4	201.6
CRT + SL [10]	14.3	22.4	114.7	193.7
CRT + CDA	**13.5**	**21.8**	**109.3**	**187.1**

TABLE IV: Comparisons with state-of-the-art methods on some shared settings, i.e., SHPartA→SHPartB, MALL→UCSD, and UCSD→MALL.

Method	Type	SHPartA → SHPartB		MALL → UCSD		UCSD → MALL	
		MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
DACC [8]	U	–	–	2.52	3.38	2.93	3.65
ASNet [11]	U	13.59	23.15	–	–	2.76	3.55
Ours	U	**12.84**	**21.92**	**2.39**	**3.26**	**2.68**	**3.50**

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 effective 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 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.

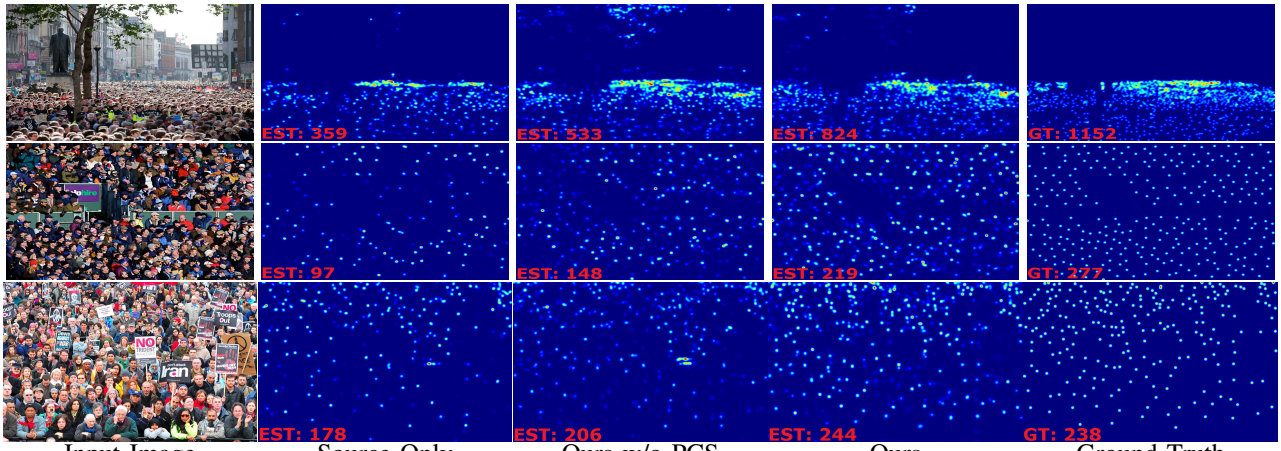

Fig. 5: Qualitative results of the estimated density maps in the Synthetic-to-Real adaptation scenario. Note that "Ours w/o PCS" also means our method without the CRT and CDA modules as PCS is the base of them.

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 Asiful Hossain, 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, Xiaoqing 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, Miaojing 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] Qiangqiang Wu, Jia Wan, and Antoni B Chan, "Dynamic momentum adaptation for zero-shot cross-domain 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 Change Loy, 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,"

in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[26] Victor Lempitsky and Andrew Zisserman, "Learning to count objects in images," in *Advances in neural information processing systems*, 2010.

[27] Zhe Lin and Larry S Davis, "Shape-based human detection and segmentation via hierarchical part-template matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 604–618, 2010.

[28] Meng Wang and Xiaogang Wang, "Automatic adaptation of a generic pedestrian detector to a specific traffic scene," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2011.

[29] Bo Wu and Ram Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, vol. 75, no. 2, pp. 247–266, 2007.

[30] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su, "Scale aggregation network for accurate and efficient crowd counting," in *European Conference on Computer Vision*, 2018.

[31] Yuhong Li, Xiaofan Zhang, and Deming 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.

[32] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu, "Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[33] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu, "Switching convolutional neural network for crowd counting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[34] Weizhe Liu, Mathieu Salzmann, and Pascal Fua, "Context-aware crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5099–5108.

[35] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, and Ling Shao, "Crowd counting and density estimation by trellis encoder-decoder networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019.

[36] Ao Luo, Fan Yang, Xin Li, Dong Nie, Zhicheng Jiao, Shangchen Zhou, and Hong Cheng, "Hybrid graph neural networks for crowd counting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[37] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang, "Attention scaling for crowd counting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020.

[38] Jia Wan and Antoni Chan, "Adaptive density map generation for crowd counting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[39] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[40] Liang Liu, Hao Lu, Hongwei Zou, Haipeng Xiong, Zhiguo Cao, and Chunhua Shen, "Weighing counts: Sequential crowd counting by reinforcement learning," in *European Conference on Computer Vision*, 2020.

[41] Viresh Ranjan, Boyu Wang, Mubarak Shah, and Minh Hoai, "Uncertainty estimation and sample selection for crowd counting," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[42] Min-hwan Oh, Peder A Olsen, and Karthikeyan Natesan Ramamurthy, "Crowd counting with decomposed uncertainty.," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[43] Vishwanath A Sindagi and Vishal M Patel, "Ha-ccn: Hierarchical attention-based crowd counting network," *IEEE Transactions on Image Processing*, vol. 29, pp. 323–335, 2019.

[44] Yunqi Miao, Zijia Lin, Guiguang Ding, and Jungong Han, "Shallow feature based dense attention network for crowd counting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[45] Anran Zhang, Jiayi Shen, Zehao Xiao, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao, "Relational attention network for crowd counting," in *Proceedings of the IEEE international conference on computer vision*, 2019.

[46] Anran Zhang, Lei Yue, Jiayi Shen, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao, "Attentional neural fields for crowd counting," in *Proceedings of the IEEE international conference on computer vision*, 2019.

[47] Dongze Lian, Jing Li, Jia Zheng, Weixin Luo, and Shenghua Gao, "Density map regression guided detection network for rgb-d crowd counting and localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019.

[48] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen, "Revisiting perspective information for efficient crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[49] Zhaoyi Yan, Yuchen Yuan, Wangmeng Zuo, Xiao Tan, Yezhen Wang, Shilei Wen, and Errui Ding, "Perspective-guided convolution networks for crowd counting," in *Proceedings of the IEEE international conference on computer vision*, 2019.

[50] Qi Zhou, Junping Zhang, Lingfu Che, Hongming Shan, and James Z Wang, "Crowd counting with limited labeling through submodular frame selection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 5, pp. 1728–1738, 2018.

[51] Mahesh Kumar Krishna Reddy, Mohammad Hossain, Mrigank Rochan, and Yang Wang, "Few-shot scene adaptive crowd counting using meta-learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020.

[52] Vishwanath A Sindagi, Rajeev Yasarla, Deepak Sam Babu, R Venkatesh Babu, and Vishal M Patel, "Learning to count in the crowd from limited labeled data," in *European Conference on Computer Vision*, 2020.

[53] Zhen Zhao, Miaojing Shi, Xiaoxiao Zhao, and Li Li, "Active crowd counting with limited supervision," in *European Conference on Computer Vision*, 2020.

[54] Yongtuo Liu, Sucheng Ren, Liangyu Chai, Hanjie Wu, Dan Xu, Jing Qin, and Shengfeng He, "Reducing spatial labeling redundancy for active semi-supervised crowd counting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[55] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe, "Weakly-supervised crowd counting learns from sorting rather than locations," in *European Conference on Computer Vision*, 2020.

[56] Zheng Xiong, Liangyu Chai, Wenxi Liu, Yongtuo Liu, Sucheng Ren, and Shengfeng He, "Glance to count: Learning to rank with anchors for weakly-supervised crowd counting," *arXiv preprint arXiv:2205.14659*, 2022.

[57] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov, "Leveraging unlabeled data for crowd counting by learning to rank," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[58] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov, "Exploiting unlabeled data in cnns by self-supervised learning to rank," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1862–1878, 2019.

[59] Nada Elassal and James H Elder, "Unsupervised crowd counting," in *Asian Conference on Computer Vision*, 2016.

[60] Deepak Babu Sam, Neeraj N Sajjan, Himanshu Maurya, and R Venkatesh Babu, "Almost unsupervised learning for dense crowd counting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[61] Mingsheng Long, Yue Cao, Zhangjie Cao, Jianmin Wang, and Michael I Jordan, "Transferable representation learning with deep adaptation networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 12, pp. 3071–3085, 2018.

[62] Baochen Sun and Kate Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European Conference on Computer Vision*, 2016.

[63] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz, "Central moment discrepancy (cmd) for domain-invariant representation learning," *arXiv preprint arXiv:1702.08811*, 2017.

[64] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua, "Residual parameter transfer for deep domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[65] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[66] Chao Chen, Zhihang Fu, Zhihong Chen, Sheng Jin, Zhaowei Cheng, Xinyu Jin, and Xian-Sheng Hua, "Homm: Higher-order moment matching for unsupervised domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[67] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, 2015.
[68] Yaroslav Ganin, Evgeniya 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.
[69] 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.
[70] 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.
[71] 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.
[72] 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.
[73] 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.
[74] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan, "Domain-symmetric networks for adversarial domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019.
[75] 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.
[76] Minghao Chen, Shuai Zhao, Haifeng Liu, and Deng Cai, "Adversarial-learned loss for domain adaptation.," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
[77] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang, "Adversarial domain adaptation with domain mixup," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
[78] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada, "Asymmetric tri-training for unsupervised domain adaptation," in *International conference on machine learning*, 2017.
[79] Zhengming Ding, Sheng Li, Ming Shao, and Yun Fu, "Graph adaptive knowledge transfer for unsupervised domain adaptation," in *European Conference on Computer Vision*, 2018.
[80] 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.
[81] 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.
[82] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.
[83] Yawei Luo, 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.
[84] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang, "Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2019.
[85] Jihan Yang, Ruijia Xu, Ruiyu 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.
[86] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi, "Differential 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.
[87] 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.
[88] 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.
[89] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick 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.
[90] 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.
[91] 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.
[92] 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.
[93] 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.
[94] Vishwanath 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.
[95] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[96] 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)-C16-S1-R]; Conv2: [K(3,3)-C16-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 [67]–[77]. 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]–[87], object detection [88]–[94].

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]:

$$\mathcal{D}(\mathbf{z}) = \sum_{k=1}^{N} \delta(\mathbf{z} - \mathbf{z}_k) * G_{\sigma_k}(\mathbf{z}), \tag{8}$$

where $\mathbf{z}$ and $\mathbf{z}_k$ denote each pixel and the k-th annotated point (total N points) in a crowd image $\mathbf{x}$. G_{σ_k} is a 2D Gaussian kernel with a bandwidth σ_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 [95] 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 with kernel size of 3×3, 64 output channels, stride size of 1, and ReLU activation function.

To measure distance between ground truth and estimated density map, we adopt the widely-used pixel-wise Euclidean loss which can be formulated as:

$$\mathcal{L}_{den}(\mathbf{x}) = \frac{1}{2M} \left\| \mathbf{D}^{est}(\mathbf{x}) - \mathbf{D}^{GT}(\mathbf{x}) \right\|_2^2 \tag{9}$$

where M is the number of pixels in the input image $\mathbf{x}$. $\mathbf{D}^{est}(\mathbf{x})$ and $\mathbf{D}^{GT}(\mathbf{x})$ represent the estimated and ground truth density maps, respectively.

2) Architecture of Point-derived Crowd Segmentation: Detailed network architecture of the Point-derived Crowd Segmentation (PCS) network is in Table VI.

TABLE VII: More ablation studies on Crowd Density Alignment (CDA) in the Synthetic-to-Real adaptation scenario.

Method	GCC → SHPartB		GCC → SHPartA	
	MAE	RMSE	MAE	RMSE
Source only	19.5	28.9	169.2	255.9
SL [10]	18.6	28.2	165.4	248.5
CDA	**16.9**	**26.7**	**160.5**	**239.6**

C. More Experiment Details

1) Implementation Details: In all experiments, we set the batch size as 2, i.e., one image per domain. We adopt random cropping and horizontal flipping for data augmentation. Adam optimizer [96] is utilized to optimize the networks with the learning rate of the crowd counter and all classifiers initialized as 10^{-4} and 10^{-5}, respectively. λ_1 and λ_2 in Eq. (6) are set to 1 and 0.3, respectively via cross validation. Following [6], scene regularization is utilized to select synthetic images from GCC to facilitate adaptation. We adopt three domain classifiers for multi-scale features extracted after each pooling layer in $G(\cdot)$ of Eq. (4). The training and evaluation are achieved on 2 NVIDIA GTX 2080Ti GPU. Evaluation metrics are MAE and RMSE [4].

2) Datasets: Six datasets are utilized in our experiments. (i) **GCC** [6] is a synthetic dataset containing 15,212 images with resolution of 1080 $\times$ 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 400 images for training and the remaining for testing. Compared to SHPartA, SHPartB has relatively fixed camera perspectives

TABLE VIII: Results of the Fixed-to-Fickle adaptation.

Fixed → Fickle (SHPartB → SHPartA)			
Method	MAE ↓	RMSE ↓	Gain ↑
Source only	194.0	298.4	–
CRT w/o PCS	153.2	247.5	21.0%/17.1%
CRT	123.3	204.6	36.4%/31.4%
CRT + CDA (Ours)	**115.6**	**199.5**	**40.4%/33.1%**
Oracle	67.5	112.1	–

TABLE IX: Results of the Normal-to-BadWeather adaptation.

Normal → BadWeather (SHPartA → JHUC)			
Method	MAE ↓	RMSE ↓	Gain ↑
Source only	208.5	535.6	–
CRT w/o PCS	173.6	437.2	16.7%/18.3%
CRT w/ PCS	159.5	394.7	23.5%/26.3%
CRT w/ PCS + CDA (full)	**153.2**	**384.0**	**26.5%/28.3%**
Oracle	80.4	215.3	–

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.