Feature-aware Adaptation and Density Alignment for Crowd Counting in Video Surveillance

Junyu Gao, *Student Member, IEEE*, Yuan Yuan, *Senior Member, IEEE*, and Qi Wang*, *Senior Member, IEEE*

***Abstract*—With the development of deep neural networks, the performance of crowd counting and pixel-wise density estimation are continually being refreshed. Despite this, there are still two challenging problems in this field: 1) current supervised learning needs a large amount of training data, but collecting and annotating them is difficult; 2) existing methods can not generalize well to the unseen domain. A recently released synthetic crowd dataset alleviates these two problems. However, the domain gap between the real-world data and synthetic images decreases the models' performance. To reduce the gap, in this paper, we propose a domain-adaptation-style crowd counting method, which can effectively adapt the model from synthetic data to the specific real-world scenes. It consists of Multi-level Feature-aware Adaptation (MFA) and Structured Density map Alignment (SDA). To be specific, MFA boosts the model to extract domain-invariant features from multiple layers. SDA guarantees the network outputs fine density maps with a reasonable distribution on the real domain. Finally, we evaluate the proposed method on four mainstream surveillance crowd datasets, Shanghai Tech Part B, WorldExpo'10, Mall and UCSD. Extensive experiments evidence that our approach outperforms the state-of-the-art methods for the same cross-domain counting problem.**

***Index Terms*—Crowd Counting, Unsupervised Domain Adaptation, Denisty Estimation.**

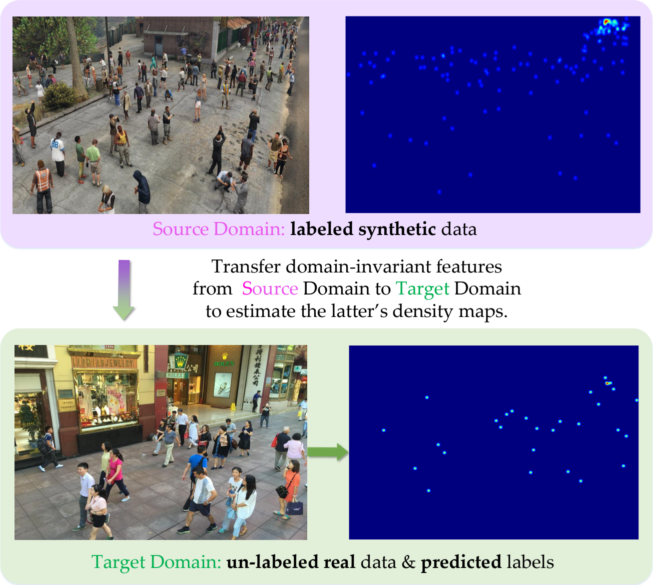

Source Domain: **labeled synthetic** data
Transfer domain-invariant features from Source Domain to Target Domain to estimate the latter's density maps.
Target Domain: **un-labeled real** data & **predicted** labels

Fig. 1: The goal of crowd counting via domain adaptation is: training a crowd counter using labeled synthetic data and applying it on the un-labeled real data.

I. Introduction

During the last half decade, Convolutional Neural Network(CNN)-based methods [1], [2], [3], [4], [5], [6], [7], [8] attain a significant progress in the field of crowd counting, especially on some research datasets including some surveillance scenes, such as UCSD [9], Mall [10], Shanghai Tech B [11], WorldExpo'10 [12] and Zhengzhou Airport [13]. Accurate counting is important to some high-level vision tasks, such as crowd analysis [14], [15], [16], [17], scene understanding [18], [19], [20] and video analysis [21], [22], [23], [24]. However, current supervised CNN-based methods need a large amount of labeled data to train a counter network. Unfortunately, the aforementioned datasets contain only a small amount of samples so that the current CNN-based methods can not be adapted to the real world. To improve the performance in the wild, theoretically, labeling more data is a potential method. Nevertheless, collecting and annotating congested crowd scenes is a laborious and expensive process. Statistically, for a still image contains $\sim$ 500 people, the labeling process takes about 30 $\sim$ 40 minutes, which is slower than other image- or object-level annotation tasks (such as image recognition [25], [26] and object detection [27] or other machine learning tasks [28]).

Considering the expensive labeling costs, some researchers focus on dealing with the scarce labeled data from two aspects: data generation and methodology. For the former, Wang *et al.* [29] construct a large-scale and synthetic GTAV Crowd Counting (GCC) dataset that automatically generated and labeled by a computer game mod. Unfortunately, the synthetic scenes are very different from the real world, of which difference is named as "domain gap". It results in a performance reduction when transferring a counter from the synthetic domain to the real-world domain.

From the perspective of methodology, Liu *et al.* [30] propose a learning-to-rank framework via leveraging unlabeled data. By this strategy, they exploit a large amount of unlabeled data to aid supervised learning. Sam *et al.* [31] present an almost unsupervised autoencoder for dense crowd counting, whose 99.9% parameters are trained without any labeled data. However, these methods still rely on manually labeled data to a different extent.

For handling the performance reduction from GCC to real data, Wang *et al.* [29], [32] propose a domain adaptation

* Q. Wang is the corresponding author.

Junyu Gao, Yuan Yuan and Qi Wang are with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: gjy3035@gmail.com; y.yuan1.ieee@gmail.com; crabwq@gmail.com).

This work was supported by the National Natural Science Foundation of China under Grant U1864204, 61773316, 61632018, and 61825603.

method via SE Cycle GAN. Fig. 1 demonstrates this problem, i.e., how to exploit free synthetic labeled data to prompt the counting performance on real-world data. SE Cycle GAN translates the synthetic image to the photo-realistic image, which is a visual and explicit adaptation. Then they train a CNN model on translated images, which performs better on real data than the CNN models without domain adaptation. However, the translated images lose some detailed information (including texture, sharpness and local structure), especially in the congested crowd regions.

Inspired by SE Cycle GAN [29], this paper focuses on domain-adaptation-style crowd counting. Different from SE Cycle GAN's explicit translation, we propose a feature-aware adaptation to reduce the domain gap at feature level and output a reasonable structured density map. It consists of a common VGG-backbone Crowd Counter (CC), Multi-level Feature-aware Adaptation (MFA) and a Structured Density map Alignment (SDA). Since the design of CC is not the core of this paper, a state-of-the-art model (SFCN [29]) is only adopted. MFA aims to extract domain-invariant features from CC, which constructs two element-wise discriminators to classify the source (from synthetic or real data) of the feature maps extracted from different layers in CC. By adversarial learning, the feature maps in CC can confuse the discriminators, so the domain gap in the feature space is effectively alleviated.

When introducing MFA, however, the quality of the density map is not good for the unseen real data. The concrete problems are: 1) some abnormal values exist in the map, 2) the map is very coarse and only reflects the density trend. To remedy these problems, we present a Structured Density map Alignment (SDA), which consists of three components. The first is similar to MFA: it is a Map Discriminator (MD) to classify the source of the density map output by CC. The second is a Self-supervised Pyramid Residual (SPR) learning on the target domain, which can maintain the consistency of density maps at different scales. The last component is named as "Map Refiner", which receives the coarse density map of CC then outputs the fine and reasonable map. Fig. 2 shows the flowchart of our proposed crowd counting via domain adaptation.

As a summary, the contributions of this paper are:

1) To our best knowledge, this paper is the first to propose a feature-aware adaptation for crowd counting.
2) This paper designs a Structured Density map Alignment (SDA) to refine the quality of density maps for unseen scenes.
3) The proposed method yields a new record of MAE and MSE on the domain-adaptation-style crowd counting from synthetic to real data.

II. Related Works

Here, we briefly review the mainstream works about the two most related tasks: crowd counting and domain adaptation from synthetic to real data.

A. Crowd Counting

1) Traditional Supervised Learning: Many algorithms are proposed to handle crowd counting in the last decade. [33], [34], [10] adopt some hand-crafted features to train the counting regressors, such as HOG, SIFT and so on. With the development of CNN, a large number of CNN-based methods [11], [35], [36], [37], [38], [39], [40] attain the significant performance. Zhang *et al.* [11] propose a multi-column CNN to cover the respective fields with different sizes for the images. To encode more contextual information, CP-CNN is presented by [35], which uses a global/local context networks to assist the counting. Considering that CP-CNN's complex training scheme, Li *et al.* [37] propose a single CNN model (CSRNet) to encode more large-range features. Cao *et al.* [38] propose a simple and efficient Scale Aggregation Network (SANet) to output structured density maps. To further explore the deep features, some methods [41], [42], [43], [44] focus on feature fusion to improve counting. In addition to directly regressing the density map, some researchers introduce other auxiliary tasks to prompt the counting performance [45], [46], [47], [48], [49], [50].

2) Counting for Scarce Labeled Data: Due to the high cost of manually labeling data, some works [51], [30], [31], [29] attempt to tackle this problems. Elassal and Elder [51] are the first to propose an unsupervised crowd counting via automatically learning how many groups of people. However, it cannot perform well in the congested crowd scenes. Liu *et al.* [30] exploit lots of unlabeled crowd scenes to reduce the estimation errors of traditional supervised learning. Sam *et al.* [31] propose an almost unsupervised autoencoder for crowd counting, of which only 0.1% parameters need labeled data during learning process. To completely get rid of manually labeled data, Wang *et al.* [29] construct a synthetic GCC dataset and propose a domain-adaptation-style method, which does not need any labeled real data. At the same time, [29] presents a novel training scheme: the model is firstly pre-trained on GCC and then fine-tuned on real-world datasets. By this strategy, the model performs better than traditional training without preliminary training on GCC.

B. Domain Adaptation

For exploiting synthetic data to prompt the classification performance on real data, some methods [52], [53], [54], [55], [56] attempt to reduce the domain gap. With the release of some synthetic segmentation datasets [57], [58], many researchers pay attention to the task of pixel-wise domain adaptation [59], [60], [61], [62], [63], [64], [65]. Hoffman *et al.* [59] firstly propose an unsupervised domain adaptation for semantic segmentation, including the global and category adaptation. Sankaranarayanan *et al.* [61] propose a joint adversarial learning approach, which transfers the target distribution to the learned embedding. Hoffman *et al.* [60] present a Cycle-Consistent Adversarial Domain Adaptation (CyCADA) for unsupervised semantic segmentation based on [66]. For producing structured segmentation masks, [67] designs a multi-level adversarial network to reduce the domain gap between synthetic and real data. To remedy the intrinsic differences in different domains, Chen *et al.* [68] propose a spatial-aware adaptation scheme to align the feature distribution of two domains. Benefiting from the [66], Chen *et al.* [64] present

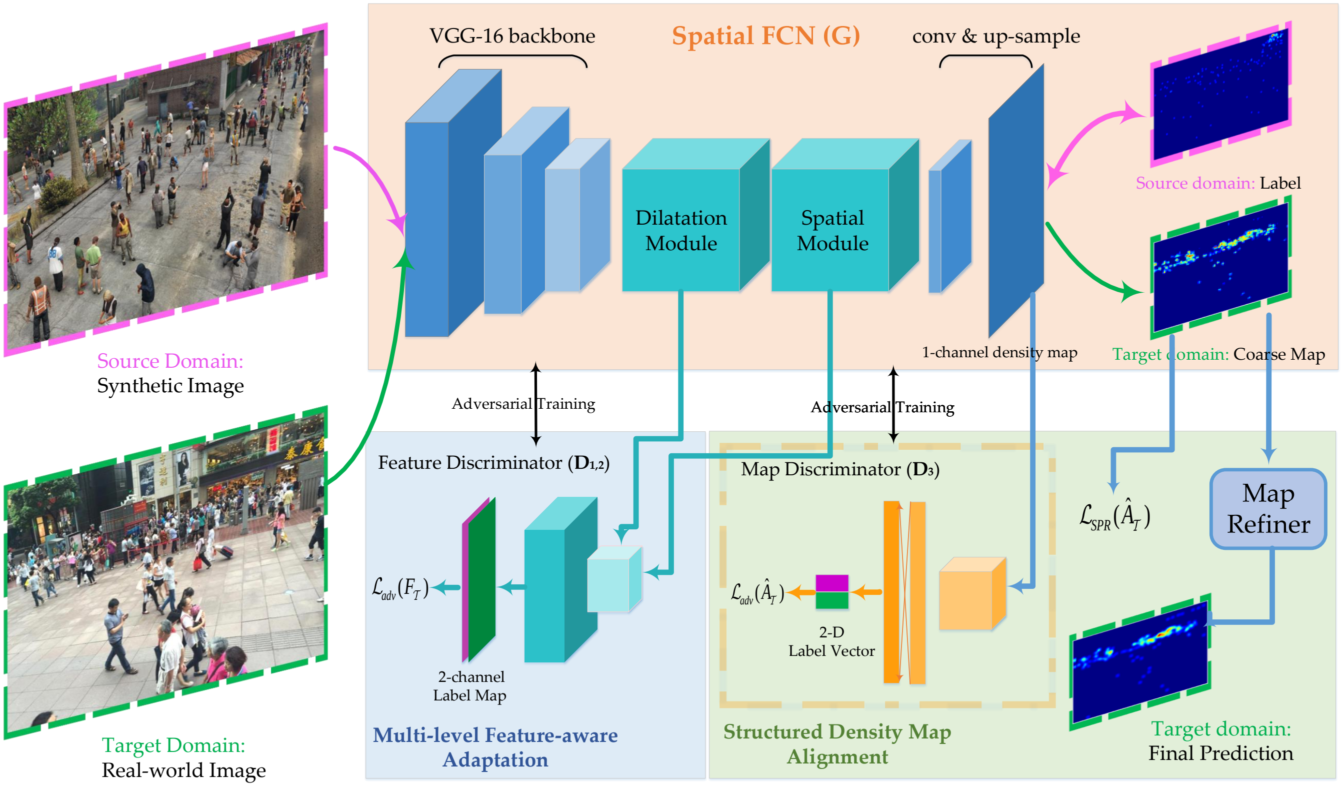

Spatial FCN (G)
VGG-16 backbone
conv & up-sample
Dilatation Module
Spatial Module
1-channel density map
Source domain: Label
Target domain: Coarse Map
Source Domain: Synthetic Image
Target Domain: Real-world Image
Adversarial Training
Adversarial Training
Feature Discriminator (D1,2)
$\mathcal{L}_{adv}(F_T)$
2-channel Label Map
Multi-level Feature-aware Adaptation
Map Discriminator (D3)
$\mathcal{L}_{adv}(\hat{A}_T)$
2-D Label Vector
$\mathcal{L}_{SPR}(\hat{A}_T)$
Map Refiner
Structured Density Map Alignment
Target domain: Final Prediction

Fig. 2: Spatial FCN (SFCN) is trained on the source domain and it directly estimates the density maps on the target domain. The bottom left box demonstrates Multi-level Feature-aware Adaptation (MFA), which classifies the pixel-wise label of feature maps. The bottom right describes Structured Density map Alignment (SDA), which consists of Map Discriminator (MD), Self-supervised Pyramid Residual (SPR) learning and Map Refiner (MR). By iteratively optimizing SFCN and the domain classifiers (MFA and MD), the final SFCN can extract the domain-invariant features.

CrDoCo for domain-adaptive dense prediction tasks, which contains two steps: 1) a CycleGAN-style image translation and 2) two task networks for the specific domain. Wang *et al.* [65] propose a two-stage density adaptation method, consisting of training on the source domain and adversarial learning on the target domain.

III. Methodology

A. Algorithmic Overview

The proposed domain-adaptation-style crowd counting consists of three modules: 1) Spatial FCN: a counter network ($\boldsymbol{G}$); 2) Multi-level Feature-aware Adaptation (MFA): two domain discriminators ($\boldsymbol{D_1}$, $\boldsymbol{D_2}$); 3) Structured Density map Alignment (SDA): a map domain discriminator ($\boldsymbol{D_3}$) and a map refiner ($\boldsymbol{R}$). The data include: source domain images $I_{\mathcal{S}}$, source domain labels $A_{\mathcal{S}}$ and target domain images $I_{\mathcal{T}}$, where $\mathcal{S}$ and $\mathcal{T}$ indicates the source and target domain, respectively.

Based on the above symbols, the goal of this paper is described as three following steps:

1) Given $I_{\mathcal{S}}$, $A_{\mathcal{S}}$ and $I_{\mathcal{T}}$, by the supervised learning on $\mathcal{S}$ for $\boldsymbol{G}$ and the adversarial learning for $\boldsymbol{G}$ and $\boldsymbol{D_i}$ $(i = 1, 2, 3)$, $\boldsymbol{G}$ can predict coarse maps $\hat{A}_{\mathcal{T}}$ of $I_{\mathcal{T}}$.
2) Using $I_{\mathcal{S}}$, $A_{\mathcal{S}}$, a synthetic counter network is trained to predict the density map $\hat{A}_{\mathcal{S}}$. Then, a map refiner $\boldsymbol{R}$ is trained using $\hat{A}_{\mathcal{S}}$ and $A_{\mathcal{S}}$, of which $\hat{A}_{\mathcal{S}}$ are inputs and $A_{\mathcal{S}}$ are labels during training process.
3) Given the coarse map $\hat{A}_{\mathcal{T}}$ in Step(1) as inputs, the trained $\boldsymbol{R}$ in Step(2) outputs final maps $\hat{A}_{\mathcal{T}}^{final}$.

Here, we formulate the loss functions during the training in Step(1) and (2). Specifically, the former is defined as:

$$\begin{aligned}\mathcal{L}\left(I_{\mathcal{S}}, A_{\mathcal{S}}, I_{\mathcal{T}}\right) =&\mathcal{L}_{cnt}(I_{\mathcal{S}}, A_{\mathcal{S}}) + \lambda\mathcal{L}_{adv}(F_{\mathcal{T}}) \\ &+ \beta\mathcal{L}_{adv}(\hat{A}_{\mathcal{T}}) + \gamma\mathcal{L}_{SPR}(\hat{A}_{\mathcal{T}}),\end{aligned} \tag{1}$$

where $\mathcal{L}_{cnt}$ is the standard MSE loss, $\mathcal{L}_{adv}$ is the adversarial loss, and $\mathcal{L}_{SPR}$ is self-supervised pyramid residual loss. λ, β and γ are the weights to balance the losses. The concrete descriptions about $\mathcal{L}_{adv}$ and $\mathcal{L}_{SPR}$ are explained in the next section. The loss function of Step(2) is the standard MSE.

B. Multi-level Feature-aware Adaptation

Since there are domain gaps between two different domains, the counter network trained by traditional supervised learning on a specific domain can not perform well on other domains. Thus, it is important that how to reduce the impact of domain gaps during the training. In other words, the purpose of adaptation is to improve the counter network to extract domain-invariant features. To this end, we present the multi-level feature-aware adaptation to reduce the domain gap of feature maps in the networks.

Since the crowd counting (density estimation) is a pixel-wise regression problem, a domain discriminator is designed to classify each unit of feature maps. To be specific, it is a fully convolutional network, including four convolution layers with leaky ReLU. It outputs a 2-channel score map with the same size as the input feature map. The values of each channel represent the confidence belonging to the source or target domain.

For the feature maps ($F_{\mathcal{S}}^i$, $F_{\mathcal{T}}^i$, $i = 1, 2$) of Dilatation and Spatial Module in $\boldsymbol{G}$, we train two discriminators for them. Through $\boldsymbol{D_i}$, the pixel-wise domain labels $O_{\mathcal{S}}^D$ and $O_{\mathcal{T}}^D$ can be obtained. For optimizing $\boldsymbol{D_i}$, we adopt 2-D pixel-wise binary cross-entropy loss as the objective function, which is formulated as:

$$\begin{aligned}\mathcal{L}_D^i(F_{\mathcal{S}}^i, F_{\mathcal{T}}^i) = &- \sum_{F_{\mathcal{S}}^i \in \mathcal{S}} \sum_{h \in H} \sum_{w \in W} \log(p(O_{\mathcal{S}}^D)) \\ &- \sum_{F_{\mathcal{T}}^i \in \mathcal{T}} \sum_{h \in H} \sum_{w \in W} \log(1 - p(O_{\mathcal{T}}^D)),\end{aligned} \tag{2}$$

where $O_{\mathcal{S}}^D$ and $O_{\mathcal{T}}^D$ are 2D-channel predicted maps, with size of $H \times W$ corresponding to the source and target inputs, and $p(\cdot)$ is a standard soft-max operation at the pixel level.

In order to extract domain-invariant features to confuse $\boldsymbol{D_i}$, the inverse adversarial loss should be added into the training process of $\boldsymbol{G}$, which is defined as:

$$\mathcal{L}_{adv}^i(F_{\mathcal{T}}^i) = -\sum_{i=1}^{2} \sum_{F_{\mathcal{T}} \in \mathcal{T}} \sum_{h \in H} \sum_{w \in W} \log(p(O_{\mathcal{T}}^D)), i = 1, 2. \tag{3}$$

This loss guides $\boldsymbol{G}$ to fool the two discriminators $\boldsymbol{D_1}$, $\boldsymbol{D_2}$. Thus, the domain gaps at different feature levels are effectively alleviated.

C. Structured Density Map Alignment

Although the domain gap between different domains is alleviated to some extent, however, there is an intractable issue to be tackled, i.e., the quality of the density map is not good for unseen data. To be specific, the map contains some abnormal regression results, and the map is coarse so that it only reflects the density trend. To this end, we propose a structured density map alignment approach to refine the coarse maps.

1) Map Discriminator: Other domain adaptation tasks [59], [60], [67] (such as image classification and segmentation) usually only focus on constructing domain discriminators at the feature level. The main reason is that: in the aforementioned tasks, each image/pixel is assigned with a specific category by soft-max or other classification layers. In other words, there is no invalid value in the prediction map. However, as for crowd counting, the regression model may produce some abnormal values on unseen domains.

To handle this problem, a map discriminator (MD) $\boldsymbol{D_3}$ is proposed to classify whether the predicted density maps from source or target domain. MD consists of three convolution layers with leaky ReLU and a fully-connected layer. It receives 1-channel prediction maps $\hat{A}_{\mathcal{S}}$ and $\hat{A}_{\mathcal{T}}$, then produces 2-D score vectors $V_{\mathcal{S}}$ and $V_{\mathcal{T}}$, which denote the confidence belonging to source or target domain. For training $\boldsymbol{D_3}$, the loss function as below is optimized:

$$\begin{aligned}\mathcal{L}_D^3(\hat{A}_{\mathcal{S}}, \hat{A}_{\mathcal{T}}) = &- \sum_{\hat{A}_{\mathcal{S}} \in \mathcal{S}} \log(p(V_{\mathcal{S}})) \\ &- \sum_{\hat{A}_{\mathcal{T}} \in \mathcal{T}} \log(1 - p(V_{\mathcal{T}})),\end{aligned} \tag{4}$$

where $p(\cdot)$ is the soft-max operation. In fact, Eq. 4 is treated as a binary cross-entropy loss. Like MFA, the inverse loss is introduced into $\boldsymbol{G}$ to fool $\boldsymbol{D_3}$, which is defined as:

$$\mathcal{L}_{adv}(\hat{A}_{\mathcal{T}}) = -\sum_{\hat{A}_{\mathcal{T}} \in \mathcal{T}} \log(p(V_{\mathcal{T}})). \tag{5}$$

2) Self-supervised Pyramid Residual Learning: Theoretically, resizing images with a similar size does not result in dramatic changes in image content. Thus, the counter is supposed to estimate the different but close number of people. However, the counter is not very robust under this circumstance. Therefore, we propose a Self-supervised Pyramid Residual (SPR) learning method to remedy the above problem.

To be specific, given a standard target image $i_{\mathcal{T}}^{1.0x} \in I_{\mathcal{T}}$, the counter $\boldsymbol{G}$ produce a density map $\hat{a}_{\mathcal{T}}^{1.0x}$. At the same time, $i_{\mathcal{T}}^{1.0x}$ is resize to a smaller and a larger image, namely $i_{\mathcal{T}}^{mx}$ $(0.8 < m < 1.0)$ and $i_{\mathcal{T}}^{nx}$ $(1.0 < n < 1.2)$, respectively. As a results, the three predicted density maps are obtained, namely $\hat{a}_{\mathcal{T}}^{1.0x}$, $\hat{a}_{\mathcal{T}}^{mx}$, and $\hat{a}_{\mathcal{T}}^{nx}$. To maintain the consistency of density map at the pixel level, we need to semantically reshape $\hat{a}_{\mathcal{T}}^{yx}$ $(y = m, n)$ to the resolution with $1.0x$. As for density maps, the semantic reshaping of $px \to qx$ $(p, q > 0)$ is defined approximately as follows:

$$a^{px} \Leftrightarrow r(a^{qx \to px}) \times \frac{q}{p} \times \frac{q}{p}, \tag{6}$$

where $r(\cdot)$ is the image resizing operation. Finally, the loss of SPR is formulated as:

$$\mathcal{L}_{SPR}(\hat{A}_{\mathcal{T}}) = \sum_{y=m,n} MSE[\hat{a}_{\mathcal{T}}^{1.0x}, p(\hat{a}_{\mathcal{T}}^{yx \to 1.0x}) \times y^2], \tag{7}$$

where $MSE[\cdot]$ is the standard Mean Squared Error.

Note that we set m and n as a value close to 1, which guarantees the image content change not much. In practice, they are randomly set in $0.8, 1.0$ and $1.0, 1.2$, respectively.

3) Map Refiner: As for the problem of coarse density maps, a Map Refiner (MR) is presented to improve the map quality. According to the MFA experiments, the producing coarse maps is very common in the cross-domain crowd counting. In order to handle this problem, we attempt to simulate the phenomenon using synthetic data and then only exploit the coarse map to reconstruct the fine map.

To be specific, in GCC dataset, the authors of [29] provide three evaluation schemes to assess the generalization ability of the model. Here, we select the cross-location splitting to mimic the cross-domain problem. In this scheme, training and testing are conducted the crowd scenes in different locations of the GTA V world, of which data are very different. Thus, the cross-location evaluation also suffers from the problem of coarse maps. After SFCN is successfully trained on GCC, it is applied to the test set and produces the prediction map. At

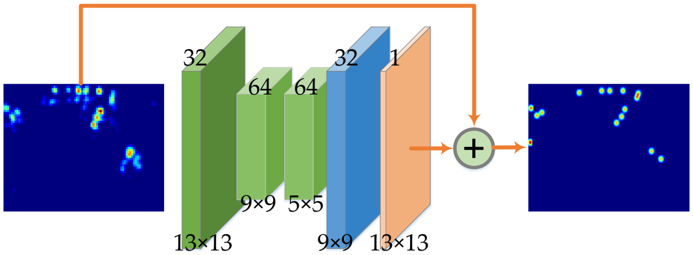


Fig. 3: The map flow of the proposed MR.

the same time, the original test data are randomly split into training/validation/testing for modeling MR according to the ratio of 70/10/20%. The MR consists of three convolutional layers, a de-convolutional layer, and a regression layer. Except for the regression layer, we adopt PReLU to restrict the output for each layer. The concrete architecture is illustrated in Fig. 3. The final output is the sum of the regression result and the input, which aims to maintain the key original density map.

In the design of MR, the large kernels (such as 13×13 and 9×9) are used to cover a larger respective field. As a result, MR can effectively reduce the low-response estimation noises and aggregate the high-response region. From the whole architecture, our goals are: 1) Design a light network to reconstruct a finer density map. So, we construct a five-layer network. 2) To maintain a high-resolution spatial feature map, we only conduct three convolutional layers and down-sample the feature map to 0.5x. Then the de-convolutional and regression layer is added to output the residual map with the size of the original input density map.

The refinement of MR is illustrated in Algorithm 1.

Algorithm 1 Algorithm for refinement pipeline of MR.

```
Input: Training data $Tr_\mathcal{S}$ and testing data $Te_\mathcal{S}$ in $\mathcal{S}$, coarse predicted maps $\hat{A}_\mathcal{T}$ in $\mathcal{T}$;
Output: The map refiner $\boldsymbol{R}$ and the refined maps $\hat{A}_\mathcal{T}^{final}$.
Train a counter network $\boldsymbol{G}_\mathcal{S}$ on $Tr_\mathcal{S}$;
Perform $\boldsymbol{G}_\mathcal{S}$ on $Te_\mathcal{S}$;
Split $Te_\mathcal{S}$ into training/validation/testing set: $Te_\mathcal{S}^{tr}$, $Te_\mathcal{S}^{val}$, $Te_\mathcal{S}^{te}$;
Train a map refiner $\boldsymbol{R}$ on $Te_\mathcal{S}^{tr}$, $Te_\mathcal{S}^{val}$ using standard MSE Loss;
Apply $\boldsymbol{R}$ on $\hat{A}_\mathcal{T}$ to obtain $\hat{A}_\mathcal{T}^{final}$;
return $\boldsymbol{R}$ and $\hat{A}_\mathcal{T}^{final}$.
```

IV. Implementation

A. Iterative optimization

During the training phase, we adopt the iterative optimization strategy to train $\boldsymbol{G}$, and $\boldsymbol{D_i}$ $(i = 1, 2, 3)$. Another network $\boldsymbol{R}$ is trained independently. The iterative optimization is explained as follows:

(1) G-update: Fix the parameter of $\boldsymbol{D_i}$ $(i = 1, 2, 3)$, and update the parameter of $\boldsymbol{G}$ by optimizing Eq. 1;

(2) D-update: Fix the parameter of $\boldsymbol{G}$, and update the parameter of $\boldsymbol{D_i}$ $(i = 1, 2, 3)$ by respectively optimizing Eq. 2 and 4.

By repeating the Step (1) and (2), the anti-domain-gap $\boldsymbol{G}$ is obtained.

The entire pipeline of training process is describe as below:

Algorithm 2 Algorithm for the training process of the entire architecture.

```
Input: The initialized Counter $\boldsymbol{G}$ and the initialized Discriminator $\boldsymbol{D_i}$ $(i = 1, 2, 3)$;
Output: The trained Counter $\boldsymbol{G}$.
repeat
    Compute the loss $\mathcal{L}_D^i$ of $\boldsymbol{D_i}$ $(i = 1, 2, 3)$ and the corresponding the gradient;
    Back-propagate gradients and update the parameter of $\boldsymbol{D_i}$ $(i = 1, 2, 3)$;
    Compute the loss $\mathcal{L}$ of $\boldsymbol{G}$ and the corresponding the gradient;
    Back-propagate gradients and update the parameter of $\boldsymbol{G}$;
until the task loss $\mathcal{L}_{cnt}$ is increased on the validation data.
return Select the best Counter $\boldsymbol{G}$ on the validation data.
```

B. Parameter Setting

This section reports the key parameter setting in the experiments. During the adversarial training, the λ, β and γ in Eq. 1 are set as 10^{-3}, 10^{-3} and 10^{-1}, respectively. The learning rates of $\boldsymbol{G}$ and $\boldsymbol{D_i}$ $(i = 1, 2, 3)$ are initialized at 10^{-5} and the $\boldsymbol{R}$'s learning rate is set as 10^{-4}. Adam [69] algorithm is performed to optimize each network and obtain the best results. The training and evaluation are performed on NVIDIA GTX 1080Ti GPU using PyTorch Crowd Counting Framework [70], [71].

C. Scene Regularization

Since GCC contains the crowd scenes under some special weathers or environments, training on the entire dataset may cause negative adaptation. To remedy the side effects, the Scene Regularization [29] is exploited to select the proper scenes. To be specific, we fully follow [29]'s setting on Shanghai Tech Part B [11] and WorldExpo'10 [12]. The selection settings on Mall [10] and UCSD [9] are described in Table I.

The Arabic numerals in the above table are explained as follows:

Level Categories 0: 0$\sim$10, 1: 0$\sim$25, 2: 0$\sim$50, 3: 0$\sim$100, 4: 0$\sim$300, 5: 0$\sim$600, 6: 0$\sim$1k, 7: 0$\sim$2k and 8: 0$\sim$4k.

Weather Categories 0: clear, 1: clouds, 2: rain, 3: foggy, 4: thunder, 5: overcast and 6: extra sunny.

Ratio range is a restriction in terms of congestion.

V. Experiments

A. Metrics

To evaluate the estimation performance of counting, the two mainstream criteria are used: Mean Absolute Error (MAE) and Mean Squared Error (MSE), which are formulated as:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|, \tag{8}$$

TABLE I: Filter condition on the four real datasets.

Target Dataset	level	time	weather	count range	ratio range
Shanghai Tech Part B	1,2,3,4,5	6:00∼19:59	0,1,5,6	10∼600	0.3∼1
WorldExpo'10	2,3,4,5,6	6:00∼18:59	0,1,5,6	0∼1000	0∼1
Mall	1,2,3,4	8:00∼18:59	0,1,5,6	0∼200	0∼1
UCSD	1,2,3,4	8:00∼18:59	0,1,5,6	0∼200	0∼1

TABLE II: The performance of No Adaptation (NoAdapt), Cycle GAN, SE Cycle GAN, FSC and our approaches on the four real-world datasets.

Method	DA	SHT B		WorldExpo'10 (MAE)						Mall		UCSD	
		MAE	MSE	S1	S2	S3	S4	S5	Avg.	MAE	MSE	MAE	MSE
NoAdpt [29]	✗	22.8	30.6	4.4	87.2	59.1	51.8	11.7	42.8	-	-	-	-
Cycle GAN [66]	✔	25.4	39.7	4.4	69.6	49.9	29.2	9.0	32.4	-	-	-	-
SE Cycle GAN [29]	✔	19.9	28.3	4.3	59.1	43.7	17.0	7.6	26.3	-	-	-	-
SE Cycle GAN (JT) [32]	✔	16.4	25.8	**4.2**	**49.6**	41.3	19.8	7.2	24.4	-	-	-	-
FSC [72]	✔	16.9	**24.7**	**4.2**	54.7	40.5	**10.5**	36.4	29.3	3.71	4.66	3.85	4.90
NoAdpt (ours)	✗	22.3	29.9	5.4	88.2	62.1	16.2	14.3	37.2	4.07	5.12	16.46	16.80
SFCN+MFA	✔	17.3	26.1	5.0	71.4	29.6	17.0	7.0	26.0	2.87	3.57	2.39	2.91
SFCN+MFA+MD+SPR	✔	16.2	24.9	6.1	60.4	23.5	16.1	**6.8**	22.6	2.56	3.48	2.09	2.51
SFCN+MFA+SDA	✔	**16.0**	**24.7**	5.7	59.9	**19.7**	14.5	8.1	**21.6**	**2.47**	**3.25**	**2.00**	**2.43**

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|^2}, \tag{9}$$

where N is the number of testing samples, y_i is the groundtruth label for counting and $\hat{y}_i$ is the estimated counting value for the i-th test sample. Besides, we adopt the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity in Image (SSIM) [73] to assess the quality of the predicted density maps.

B. Datasets

For evaluating the proposed method, we conduct the adaptation experiments from GCC Dataset [29] to another four real-world datasets containing the consistent crowd scene, namely Shanghai Tech Part B [11], WorldExpo'10 [12], Mall [10] and UCSD [9].

GTA V Crowd Counting Dataset (GCC), a large-scale synthetic dataset, consists of still $15,212$ crowd images with resolution of 1080×1920, which are captured from 400 surveillance cameras in 100 locations of a fictional city.

Shanghai Tech Part B [11] is collected from the surveillance camera on the Nanjing Road Pedestrian Street in Shanghai, China. It contains 400 training and 316 testing images with the same resolution of 768×1024. The entire dataset contains $88,488$ pedestrians.

WorldExpo'10 is a cross-scene large-scale crowd counting dataset, which is present by Zhang *et al.* [12]. All images are captured from 108 surveillance cameras in Shanghai 2010 WorldExpo, of which images from the 103 cameras are training data and the others are testing data. To be specific, WorldExpo'10 contains $3,980$ images with size of 576×720 and $199,923$ labeled pedestrians.

Mall [10] is an indoor crowd counting dataset, which is collected using a surveillance camera installed in a shopping mall. The dataset records the 2,000 sequential frames with resolution of 480×640. The first 800 frames are training samples and the others are test samples.

UCSD [9] is a single-scene dataset collected from a video camera at a pedestrian walkway. The dataset contains $2,000$ frames with a low resolution of 158×238, of which 601 to 1,400 images are training data and the others are testing samples.

C. Adaptation to Real-world Datasets

This section compares the proposed methods with other mainstream crowd counting via domain adaptation. All methods adopt Spatial FCN (SFCN) as a counter. Cycle GAN [66] and SE Cycle GAN [29] translate synthetic scenes to photo-realistic scenes, then they train a counter network on translated images. Finally, the counter is applied to real-world datasets. Table II lists the performance of No Adaptation (NoAdapt), Cycle GAN [66], SE Cycle GAN [29], FSC[72] and our approaches on the four real-world datasets. In order to show the real improvement, we also re-implement the SFCN without adaptation (NoAdpt) using our code, of which results are close to NoAdpt by [29]. Note that the results on WorldExpo'10 only report MAE and the final performance is the average of MAEs on five scenes. From the entire table, our proposed full model (SFCN+MFA+SDA, namely SFCN+MFA+MD+SPR+MR) outperforms other state-of-the-art methods on all datasets. Notably, compared with FSC that uses GAN for feature maps in one layer, the proposed method adopts GAN for multiple layer, which makes model extract more domain-invariant features.

Take the results of Shanghai Tech B as an example, MFA decreases the 22.3% MAE and the 12.7% MSE, which implies the proposed MFA can effectively reduce the domain gap between synthetic and real-world data. When introducing

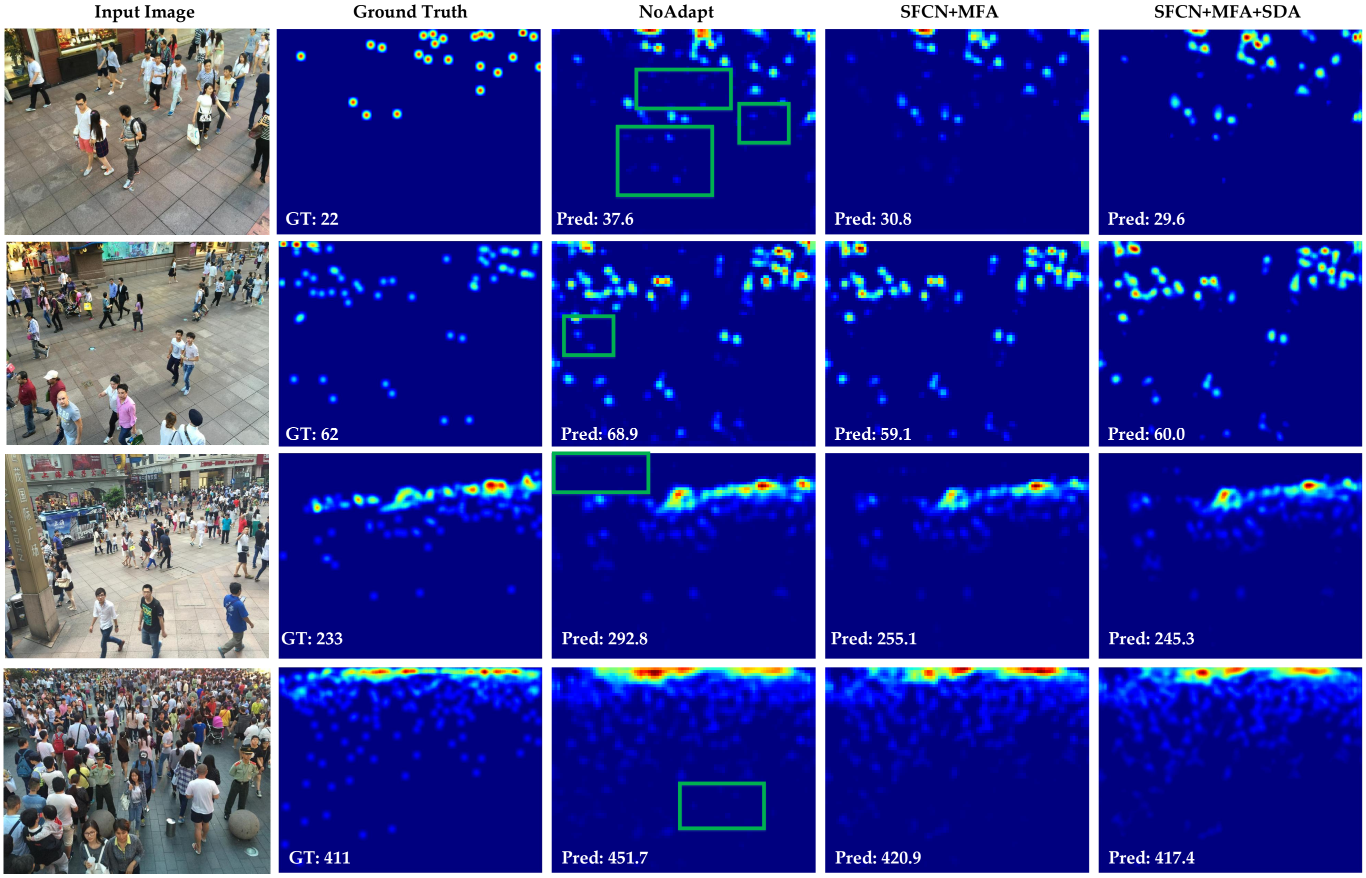


Fig. 4: Exemplar results of adaptation from GCC to Shanghai Tech Part B dataset. From left to right: input image, ground truth, and the predictions of NoAdpt, SFCN+MFA, and SFCN+MFA+SDA.

SDA, the errors are further alleviated. As a result, our proposed method achieves the MAE of 16.4 and the MSE of 25.4, which is better than the result of SE Cycle GAN (MAE/MSE: 19.9/28.3).

Fig. 4 shows the visualization results on Shanghai Tech Part B dataset. From Column 3, NoAdapt produces some estimation errors on the background, especially in the green boxes. When adopting MFA and SDA, the aforementioned errors are effectively reduced. By comparing Column 4 and 5, we find that SDA yields more reasonable and finer density maps (recommend readers to zoom in the image for comparison). As a result, the counting values of Column 5 are closer to the ground truth than that of Column 4.

D. Ablation Study on Shanghai Tech B

Analysis of different modules. Table III reports the results of our proposed step-wise methods: No Adaptation (NoAdpt), SFCN+MFA, SFCN+SDA and SFCN+MFA+SDA. The error reductions of SDA are smaller than those of MFA. Compared with NoAdpt, MFA decreases 22.3% in MAE and 12.7% in MSE, but SDA only reduces 11.7/5.0% in terms of MAE/MSE. The main reason is that SDA mainly aims to improve the quality of maps, but MFA focuses on aiding SFCN to extract domain-invariant features, which is the core of reducing the domain gap.

Comparison with Cycle-GAN-style methods. Here, we compare the quality of density maps of our proposed algorithms with Cycle-GAN-style methods [66], [60], [29] on Shanghai Tech B in Table III. All methods adopt the same SFCN with VGG-16 backbone as the counter network. As for the four metrics, the proposed adaptation algorithms perform better than Cycle-GAN-style methods. In terms of the methodology, GAN and SFCN are trained separately in Cycle-GAN-style methods: 1) train GAN and translate the image; 2) train SFCN using translated images. This separated training scheme can not attain a good result. In our method, the training of SFCN and adversarial networks is applied iteratively, which guarantees the entire models achieve better performance.

Comparison with feature-level adversarial learning methods. Compared with FCN Wld w/o CA [59] and CODA [65] that also attempt to extract domain-invariant features in a specific layer of the networks, the proposed MFA outperforms them (MAE: 17.3 (MFA) *vs.* 23.1 [59] and 19.2 [65]) due to it conducts adversarial learning on the high-level semantic features and multiple layers, which causes that the model extracts more effective domain-invariant features. In the experiment, note that FCN Wld contains category adaptation (CA) for pixel-wise classification. Thus, we remove CA from FCN Wld. At the same time, we replace the classification layer with a regression layer in FCN Wld. Correspondingly, the task loss is MSE loss instead of Cross Entropy Loss.

TABLE III: The performance on Shanghai Tech Part B.

Methods	Src data	MAE	MSE	PSNR	SSIM
NoAdpt [29]	GCC	22.8	30.6	24.66	0.715
FCN Wld w/o CA [59]	GCC	23.1	29.4	24.12	0.732
Cycle GAN [66]	GCC	25.4	39.7	24.60	0.763
CyCADA [60]	GCC	18.7	26.6	25.58	0.826
SE Cycle GAN [29]	GCC	19.9	28.3	24.78	0.765
CODA [65]	GCC	19.2	28.5	24.63	0.803
NoAdpt (ours)	GCC	22.3	29.9	25.02	0.811
SFCN+MFA	GCC	17.3	26.1	25.37	0.837
SFCN+SDA	GCC	19.7	28.4	25.11	0.832
SFCN+MFA+MD+SPR	GCC	16.2	24.9	25.45	0.846
SFCN+MFA+SDA	GCC	16.0	24.7	25.62	**0.856**
NoAdpt [65]	SHT A	27.3	36.2	-	-
CODA [65]	SHT A	15.9	26.9	-	-
SFCN+MFA+SDA	SHT A	**14.8**	**21.9**	**25.89**	0.844

Effect of Map Refiner. In our proposed method, Map Refiner (MR) is an independent network, which is trained on the testing set on GCC dataset. It can effectively refine the quality of SFCN's prediction maps. Table IV lists the quantitative results with/without MR. From it, we indeed find MR prompts the performance of counting and density map quality. In order to intuitively show the effect of MR, the density maps are compared in Fig. 5. According to the comparison, there are three main effects of MR: 1) MR produces finer density maps for the entire outputs; 2) MR yields more independent density region, such as the green box; 3) MR reduces the estimation errors for low-response regions, such as the red box. In general, the proposed MR outputs the more reasonable and finer density maps.

TABLE IV: The effect of each module in SDA on ShanghaiTech B.

Methods	MAE	MSE	PSNR	SSIM
SFCN+MFA+MD	16.5	25.8	25.41	0.845
SFCN+MFA+MD+SPR	16.2	24.9	25.45	0.846
SFCN+MFA+MD+SPR+MR	**16.0**	**24.7**	**25.62**	**0.856**

VI. Analysis and Discussion

A. Selection of Feature Maps in MFA

In MFA, the feature maps of Dilatation and Spatial Module are extracted to train the two domain discriminators. In fact, there are many potential feature maps in SFCN that are selected as the inputs for domain discriminators. In this section, we analyze the effects under different combinations of different layers' outputs on experimental results. To be specific, three types of feature maps are selected, namely the outputs of conv4_3, Dilatation Module and Spatial Module.

Table V reports the estimation errors under different settings on Shanghai Tech Part B. From the results of single-level adversarial experiments, the domain gap can be reduced more effectively on the deep layer (after Dilation and Spatial) than the shallow layer (conv4_3). In the double-level experiments, the combination of Dilation and Spatial Module is the best, which achieves the MAE of 17.3 and the MSE of 26.1.

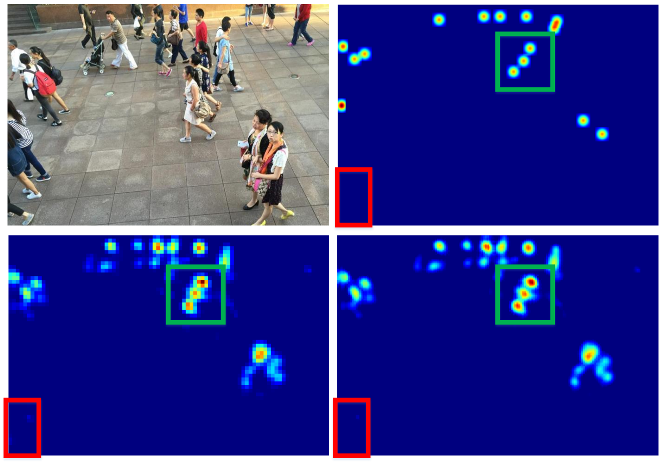


Fig. 5: The comparison of with/without MR. Row 1: original image, ground truth; Row 2: without MR, with MR.

TABLE V: The comparison results of different combinations of feature maps.

Methods	Combination			Errors	
	conv4_3	Dilation	Spatial	MAE	MSE
NoAdapt				22.3	29.9
single	✔			20.3	28.1
		✔		18.5	27.2
			✔	18.4	29.1
double	✔	✔		18.6	27.8
	✔		✔	19.3	29.7
		✔	✔	**17.3**	**26.1**
triple	✔	✔	✔	**17.3**	26.3

When adopting triple-level adversarial learning, the results are very close to that of Dilation+Spatial. We think the deep-layer adversarial training can reduce the domain gap to the maximum extent. It is not necessary that introducing shallow-layer adversarial learning. Therefore, Dilation+Spatial is the final combination in MFA.

Furthermore, we discuss why Dilation+Spatial is the best combination. The main reasons are: 1) the two modules encode more large-range contextual information than that of the backbone. 2) crowd density is a type of high-level features. For a deep CNN, shallow layers only extract low-level features, such as texture, edge or other appearance characteristics. Deep layers can extract high-level semantic features. In the two data domains, difference of low-level features of is an objective reality. But the crowd density features are the same. Thus, we select Dilation and Spatial Module to construct adversarial learning, which makes models to learn domain-invariant features.

B. Adaptation between Real-world Datasets

The domain gap does not only exist in the adaptation between the synthetic and the real world but also in the transferring process of different real scenes. For example, real-world datasets are very different in terms of scene attributes, cameras, and so on. In this section, we select two typical real-world datasets (Mall and UCSD) to evaluate the proposed method. These two datasets are very different: Mall is an indoor crowd dataset captured by RGB cameras but UCSD is an outdoor crowd dataset captured by gray-scale sensors. Table VI reports the results of two different adaptation experiments (Mall→UCSD and UCSD→Mall). From it, the proposed method effectively reduces the domain gap between different datasets and attains an acceptable counting result, which shows that our method is important for landing the counter network in real life.

TABLE VI: The performance of adaptation between the two real-world datasets.

Methods	Mall→UCSD		UCSD→Mall	
	MAE	MSE	MAE	MSE
NoAdpt	15.80	16.14	3.01	3.74
CODA [65]	-	-	3.38	4.15
SFCN+MFA+SDA	**2.08**	**2.58**	**2.66**	**3.32**

C. Generalization Ability on Other Models

In this paper, we conduct the adaptation experiments on the SFCN. In fact, the proposed method can be applied to any FCN-based crowd counter. Here, we adopt two classical crowd counters, MCNN [11] and CSRNet [37], to verify the proposed adaption method on Shanghai Tech Part B Dataset. For CSRNet, we select the feature maps of the VGG-16 backbone and dilated convolutional layers as the input for Multi-level Feature-aware Adaptation (MFA). Since MCNN only contains single-stage convolution, we have to apply Single-level Feature-aware Adaptation (SFA) on MCNN.

TABLE VII: The comparison of different crowd counter in the proposed adaptation method.

Methods	NoAdpt		Adapted	
	MAE	MSE	MAE	MSE
MCNN	66.1	82.8	28.0	43.9
CSRNet	23.0	33.9	18.1	26.9
SFCN	**22.3**	**29.9**	**16.4**	**25.4**

Table VII lists the comparison results on the three different crowd counters. From it, we find the estimation errors are significantly decreased after adaptation. This phenomenon indicates our method can be generalized to other CNN-based crowd counters to reduce the domain gap.

D. Adaptation to Inconsistent Real Data

Usually, the same surveillance cameras are equipped in a city or a specific region by the government. As a result, the captured scenes are highly consistent. Considering that, this work focuses on the domain adaptation to the consistent crowd scenes, which is practical in the real scene application. For the traditional crowd counting, there are two real-world datasets that contains inconsistent crowd scenes collected from Internet, namely, Shanghai Tech Part A [11] and UCF-QNRF [41], of which images are captured by people instead of surveillance cameras. Here, we verify the proposed method on these datasets.

TABLE VIII: The performance of No Adaptation (NoAdapt), Cycle GAN[66], SE Cycle GAN [29] and our approaches on Shanghai Tech Part A and UCF-QNRF dataset.

Method	DA	SHT A		UCF-QNRF	
		MAE	MSE	MAE	MSE
NoAdpt [29]	✗	160.0	216.5	275.5	458.5
Cycle GAN [66]	✔	143.3	204.3	257.3	400.6
SE Cycle GAN [29]	✔	**123.4**	**193.4**	**230.4**	**384.5**
NoAdpt (ours)	✗	156.4	210.7	269.5	480.2
SFCN+MFA+SDA	✔	144.6	200.6	255.4	407.9

The results on these two datasets are shown in Table VIII. From it, we find the results of the proposed method is far from that of SE Cycle GAN [29]. The main reason is that inconsistent real data causes that the GANs ($\boldsymbol{G}$ and $\boldsymbol{D_i}$ ($i = 1, 2, 3$)) are hard to converge. Especially, the loss for $\boldsymbol{G}$ suffers from severe fluctuations.

In summary, our method is more suitable for adaptation from synthetic data to real-world surveillance scenes.

E. The Effect of Large Kernel in MR

In Section III-C3, we design a Map Refiner with large kernel size to cover a large respective field. Here, we conduct some experiments to verify our network design. To be specific, three Map Refiners are designed with different settings for the number of kernel sizes: $[13, 9, 5, 9, 13]$, $[9, 5, 3, 5, 9]$ and $[7, 5, 3, 5, 7]$ (Fig. 3's setting is the first one). Table IX reports the refiner results in terms of MAE, MSE, PSNR and SSIM. From it, we find the setting of $[13, 9, 5, 9, 13]$ obtains the best refinement performance. The main reason is that large kernel cover the large-range spatial information to reduce the low-response noises and improve the reconstruction performance for a whole predicted head density region.

TABLE IX: The comparison of different kernel sizes in the Map Refiner on ShanghaiTech Part B.

Settings	MAE	MSE	PSNR	SSIM
None	16.24	24.91	25.45	0.846
$[13, 9, 5, 9, 13]$	16.02	24.73	25.62	0.856
$[9, 5, 3, 5, 9]$	16.09	24.81	25.58	0.849
$[7, 5, 3, 5, 7]$	16.20	24.87	25.47	0.841

F. Introducing SPR into Classical Supervised Models

In this section, we implement SPR in the training of MCNN and CSRNet on ShanghaiTech Part B dataset. Table X illustrates the improvements after introducing the SPR in the Traditional Supervision (TS). From the table, we find the estimation errors are reduced significantly. Notably, take MCNN as an example, MAE/MSE decreases from 23.3/40.2 to 21.8/37.7 (relatively 6.4/6.2%) on ShanghaiTech Part B.

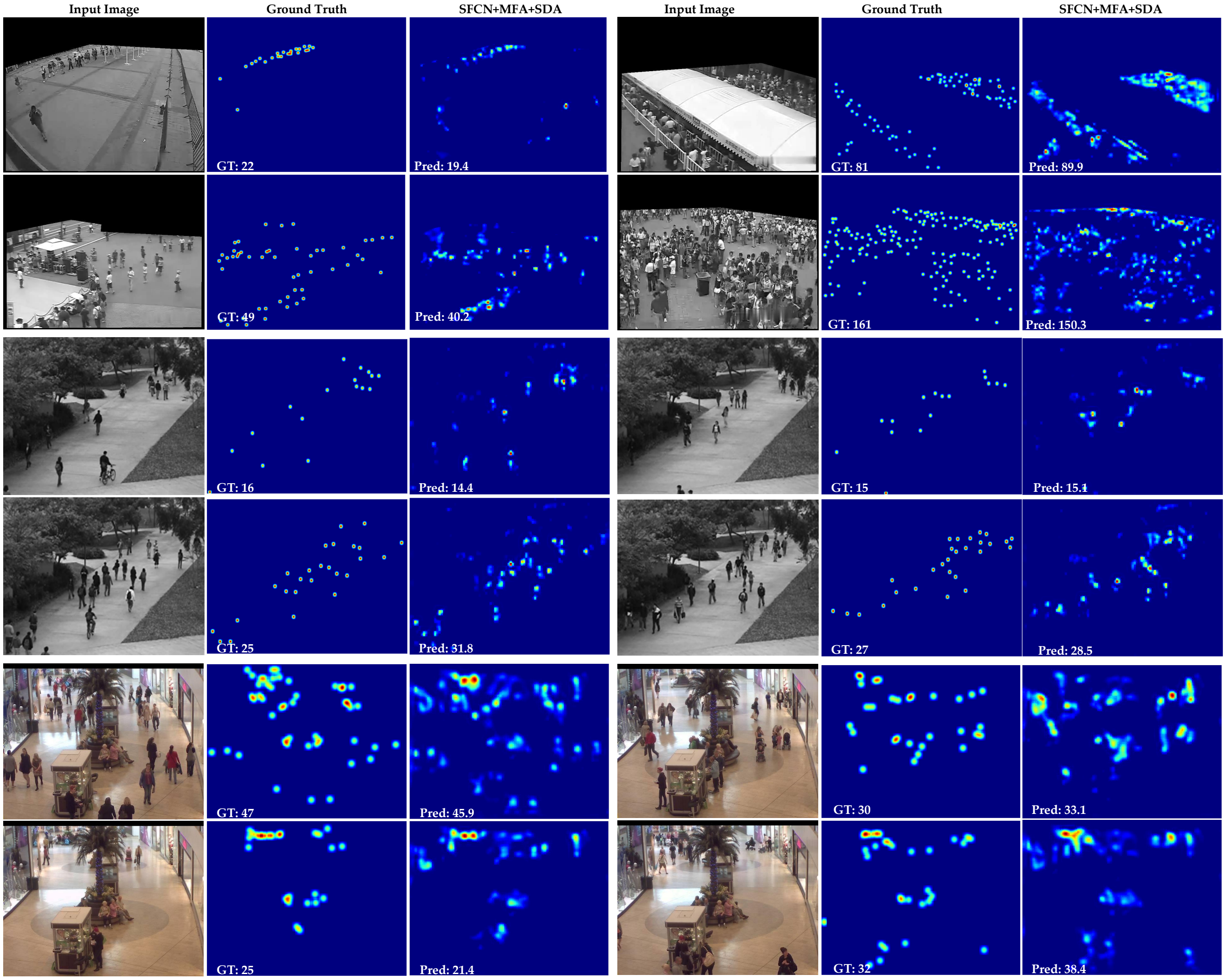


Fig. 6: Exemplar results of adaptation from GCC to WorldExpo’10, UCSD, and Mall dataset.

In fact, SPR can be treated as a combination of multi-scale training and self-supervised algorithms, which makes the model more robust.

TABLE X: The effect of SPR in traditional supervised models on ShanghaiTech Part B.

Method	TS	TS+SPR
MCNN	23.3/40.2	21.8/37.7
CSRNet	10.6/17.1	10.1/15.4

G. Visualization Results

This section demonstrates the visualization results on WorldExpo’10 [12], Mall [10] and UCSD [9] datasets in Fig. 6. In general, the adaptation results are able to reflect the crowd density and predict the number of people approximately. However, compared with ground truth, there are many error estimations in the background region, especially in some places far away from the camera. The main reason is these places are very similar to head in terms of texture. Note that there are some shifts in the visualization results on the UCSD dataset. The main reason is that the key-point location in UCSD is the center of the person, but the key-point location in GCC is the center of the head.

VII. Conclusion

In this work, we present a GAN-based adaptation method for crowd counting by learning from synthetic data and the corresponding free labels. The proposed method consists of two modules: Multi-level Feature-aware Adaptation (MFA) and a Structured Density map Alignment (SDA). The former module focuses on reducing the domain gap between synthetic and real data at the feature level, which is the first attempt in crowd counting. The latter aims to produce reasonable and fine density maps on the unseen data. Experimental results show that the proposed method outperforms the state-of-the-art performance for the same task. For the future work, since the high-level semantic information (such as the structured features of persons or groups) are more robust and invariant than pixel-level features in the cross-domain problem, we will attempt to introduce these features into domain adaptation to prompt the counting performance in the real world.

References

[1] D. Onoro-Rubio and R. J. López-Sastre, “Towards perspective-free object counting with deep learning,” in *European Conference on Computer Vision*. Springer, 2016, pp. 615–629.

[2] Y. Hu, H. Chang, F. Nian, Y. Wang, and T. Li, “Dense crowd counting from still images with convolutional neural networks,” *Journal of Visual Communication and Image Representation*, vol. 38, pp. 530–539, 2016.

[3] D. Babu Sam, S. 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, pp. 5744–5752.

[4] D. Babu Sam, N. N. Sajjan, R. Venkatesh Babu, and M. Srinivasan, “Divide and grow: capturing huge diversity in crowd images with incrementally growing cnn,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3618–3626.

[5] M. Shi, Z. Yang, C. Xu, and Q. Chen, “Revisiting perspective information for efficient crowd counting,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7279–7288.

[6] J. Wan, W. Luo, B. Wu, A. B. Chan, and W. Liu, “Residual regression with semantic prior for crowd counting,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4036–4045.

[7] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin, “Crowd counting with deep structured scale integration network,” *arXiv preprint arXiv:1908.08692*, 2019.

[8] Y. Zhou, J. Yang, H. Li, T. Cao, and S. Kung, “Adversarial learning for multiscale crowd counting under complex scenes,” *IEEE Transactions on Cybernetics*, pp. 1–10, 2020.

[9] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, “Privacy preserving crowd monitoring: Counting people without people models or tracking,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–7.

[10] K. Chen, C. C. Loy, S. Gong, and T. Xiang, “Feature mining for localised crowd counting.” in *BMVC*, vol. 1, no. 2, 2012, p. 3.

[11] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. 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.

[12] C. Zhang, K. Kang, H. Li, X. Wang, R. Xie, and X. Yang, “Data-driven crowd understanding: a baseline for a large-scale crowd dataset,” *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1048–1061, 2016.

[13] X. Jiang, L. Zhang, P. Lv, Y. Guo, R. Zhu, Y. Li, Y. Pang, X. Li, B. Zhou, and M. Xu, “Learning multi-level density maps for crowd counting,” *IEEE transactions on neural networks and learning systems*, 2019.

[14] V. J. Kok and C. S. Chan, “Grcs: Granular computing-based crowd segmentation,” *IEEE Transactions on Cybernetics*, vol. 47, no. 5, pp. 1157–1168, May 2017.

[15] A. S. Rao, J. Gubbi, S. Marusic, and M. Palaniswami, “Crowd event detection on optical flow manifolds,” *IEEE Transactions on Cybernetics*, vol. 46, no. 7, pp. 1524–1537, 2016.

[16] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, “Crowded scene analysis: A survey,” *IEEE transactions on circuits and systems for video technology*, vol. 25, no. 3, pp. 367–386, 2014.

[17] Y. Yuan, Y. Feng, and X. Lu, “Structured dictionary learning for abnormal event detection in crowded scenes,” *Pattern Recognition*, vol. 73, pp. 99–110, 2018.

[18] K. K. Santhosh, D. P. Dogra, P. P. Roy, and B. B. Chaudhuri, “Trajectory-based scene understanding using dirichlet process mixture model,” *IEEE Transactions on Cybernetics*, pp. 1–14, 2019.

[19] B. Zhao, X. Li, and X. Lu, “Property-constrained dual learning for video summarization,” *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[20] Z. Xiong, Y. Yuan, N. Guo, and Q. Wang, “Variational context-deformable convnets for indoor scene parsing,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 3991–4001.

[21] Y. Yuan, Y. Lu, and Q. Wang, “Tracking as a whole: Multi-target tracking by modeling group behavior with sequential detection,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 12, pp. 3339–3349, 2017.

[22] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, and X. Zhou, “Semisupervised feature selection via spline regression for video semantic recognition,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 2, pp. 252–264, 2015.

[23] B. Zhao, X. Li, and X. Lu, “Cam-rnn: Co-attention model based rnn for video captioning,” *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5552–5565, 2019.

[24] Z. Xiong, Y. Yuan, and Q. Wang, “MSN: modality separation networks for RGB-D scene recognition,” *Neurocomputing*, vol. 373, pp. 81–89, 2020.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[26] B. Zhao, L. Hua, X. Li, X. Lu, and Z. Wang, “Weather recognition via classification labels and weather-cue maps,” *Pattern Recognition*, vol. 95, pp. 272–284, 2019.

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[28] Z. Wang, F. Nie, L. Tian, R. Wang, and X. Li, “Discriminative feature selection via a structured sparse subspace learning module,” in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 2020, pp. 3009–3015.

[29] Q. Wang, J. Gao, W. Lin, and Y. Yuan, “Learning from synthetic data for crowd counting in the wild,” *arXiv preprint arXiv:1903.03303*, 2019.

[30] X. Liu, J. van de Weijer, and A. D. Bagdanov, “Leveraging unlabeled data for crowd counting by learning to rank,” *arXiv preprint arXiv:1803.03095*, 2018.

[31] D. B. Sam, N. N. Sajjan, H. Maurya, and R. V. Babu, “Almost unsupervised learning for dense crowd counting,” 2019.

[32] Q. Wang, J. Gao, W. Lin, and Y. Yuan, “Pixel-wise crowd understanding via synthetic data,” *International Journal of Computer Vision*, pp. 1–21, 2020.

[33] D. Kong, D. Gray, and H. Tao, “A viewpoint invariant approach for crowd counting,” in *18th International Conference on Pattern Recognition (ICPR’06)*, vol. 3. IEEE, 2006, pp. 1187–1190.

[34] V. Rabaud and S. Belongie, “Counting crowded moving objects,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, vol. 1. IEEE, 2006, pp. 705–711.

[35] V. A. Sindagi and V. M. Patel, “Generating high-quality crowd density maps using contextual pyramid cnns,” in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 1879–1888.

[36] D. B. Sam and R. V. Babu, “Top-down feedback for crowd counting convolutional neural network,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[37] Y. Li, X. Zhang, and D. Chen, “Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1091–1100.

[38] X. Cao, Z. Wang, Y. Zhao, and F. Su, “Scale aggregation network for accurate and efficient crowd counting,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[39] V. Ranjan, H. Le, and M. Hoai, “Iterative crowd counting,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 270–285.

[40] J. Gao, Q. Wang, and Y. Yuan, “Scar: Spatial-/channel-wise attention regression networks for crowd counting,” *Neurocomputing*, vol. 363, pp. 1–8, 2019.

[41] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, “Composition loss for counting, density map estimation and localization in dense crowds,” *arXiv preprint arXiv:1808.01050*, 2018.

[42] W. Liu, M. Salzmann, and P. Fua, “Context-aware crowd counting,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5099–5108.

[43] V. A. Sindagi and V. M. Patel, “Multi-level bottom-top and top-bottom feature fusion for crowd counting,” *arXiv preprint arXiv:1908.10937*, 2019.

[44] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, and L. Shao, “Crowd counting and density estimation by trellis encoder-decoder networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6133–6142.

[45] V. A. Sindagi and V. M. Patel, “Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting,” in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2017, pp. 1–6.

[46] J. Gao, Q. Wang, and X. Li, "Pcc net: Perspective crowd counting via spatial convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[47] M. Zhao, J. Zhang, C. Zhang, and W. Zhang, "Leveraging heterogeneous auxiliary tasks to assist crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 736–12 745.

[48] G. He, Z. Ma, B. Huang, B. Sheng, and Y. Yuan, "Dynamic region division for adaptive learning pedestrian counting," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1120–1125.

[49] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, "Density map regression guided detection network for rgb-d crowd counting and localization," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[50] J. Wan, N. S. Kumar, and A. B. Chan, "Fine-grained crowd counting," *arXiv preprint arXiv:2007.06146*, 2020.

[51] N. Elassal and J. H. Elder, "Unsupervised crowd counting," pp. 329–345, 2016.

[52] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[53] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2962–2971.

[54] J. Wen, R. Liu, N. Zheng, Q. Zheng, Z. Gong, and J. Yuan, "Exploiting local feature patterns for unsupervised domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5401–5408.

[55] J. Wen, N. Zheng, J. Yuan, Z. Gong, and C. Chen, "Bayesian uncertainty matching for unsupervised domain adaptation," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 3849–3855.

[56] Q. Wang, J. Gao, W. Lin, and X. Li, "Nwpu-crowd: A large-scale benchmark for crowd counting and localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[57] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *European Conference on Computer Vision*. Springer, 2016, pp. 102–118.

[58] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3234–3243.

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

[60] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," *arXiv preprint arXiv:1711.03213*, 2017.

[61] S. Sankaranarayanan, Y. Balaji, A. Jain, S. Nam Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3752–3761.

[62] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4376–4386, 2019.

[63] H. Lee, H. Tseng, J. Huang, M. Singh, and M. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 36–52.

[64] Y. Chen, Y. Lin, M. Yang, and J. Huang, "Crdoco: Pixel-level domain transfer with cross-domain consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1791–1800.

[65] W. Li, L. Yongbo, and X. Xiangyang, "Coda: Counting objects via scale-aware adversarial density adaption," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 193–198.

[66] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.

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

[68] Y. Chen, W. Li, and L. Van Gool, "Road: Reality oriented adaptation for semantic segmentation of urban scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7892–7901.

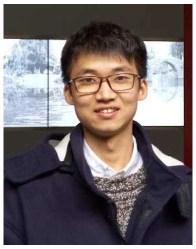
[69] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[70] J. Gao, W. Lin, B. Zhao, D. Wang, C. Gao, and J. Wen, "C^3 framework: An open-source pytorch code for crowd counting," *arXiv preprint arXiv:1907.02724*, 2019.

[71] B. Steiner, Z. Devito, S. Chintala, S. Gross, A. Paszke, F. Massa, A. Lerer, G. Chanan, Z. Lin, E. Yang *et al.*, "Pytorch: An imperative style, high-performance deep learning library," pp. 8026–8037, 2019.

[72] T. Han, J. Gao, Y. Yuan, and Q. Wang, "Focus on semantic consistency for cross-domain crowd understanding," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1848–1852.

[73] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.



Junyu Gao received the B.E. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China, in 2015. He is currently pursuing the Ph.D. degree from Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Yuan Yuan (M'05-SM'09) is currently a Full Professor with the School of Computer Science and the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science, with the Unmanned System Research Institute, and with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.