Domain-adaptive Crowd Counting via High-quality Image Translation and Density Reconstruction

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

***Abstract*—Recently, crowd counting using supervised learning achieves a remarkable improvement. Nevertheless, most counters rely on a large amount of manually labeled data. With the release of synthetic crowd data, a potential alternative is transferring knowledge from them to real data without any manual label. However, there is no method to effectively suppress domain gaps and output elaborate density maps during the transferring. To remedy the above problems, this paper proposes a Domain-Adaptive Crowd Counting (DACC) framework, which consists of a high-quality image translation and density map reconstruction. To be specific, the former focuses on translating synthetic data to realistic images, which prompt the translation quality by segregating domain-shared/independent features and designing content-aware consistency loss. The latter aims at generating pseudo labels on real scenes to improve the prediction quality. Next, we retrain a final counter using these pseudo labels. Adaptation experiments on six real-world datasets demonstrate that the proposed method outperforms the state-of-the-art methods.**

***Index Terms*—Crowd Counting, Domain Adaptation, Image Translation**

I. Introduction

Crowd counting is usually treated as a pixel-level estimation problem, which predicts the density value for each pixel and sums the entire prediction map as a final counting result. A pixel-wise density map produces more detailed information than a single number for a complex crowd scene. In addition, it also boosts other highly semantic crowd analysis (group detection [1], [2], [3], crowd segmentation [4], public management [5], *etc.*) or video surveillance tasks (video summarization [6], [7], [8] and abnormal detection [9]). Recently, benefiting from the powerful capacity of deep learning, there is a significant promotion in the field of counting. However, currently released datasets are too small to satisfy the mainstream deep-learning-based methods [10], [11], [12], [13], [14], [15]. The main reason is that constructing a large-scale crowd counting dataset is extremely demanding, which needs many human resources [16].

To handle the scarce data problem, many researchers pay attentions to data generation. Exploiting computer graphics to render photo-realistic crowd scenes becomes an alternative to generate a large-scale dataset [17]. Unfortunately, due to the differences between the synthetic and real worlds (also named as "domain shifts/gap"), there is an obvious performance degradation when applying the synthetic crowd model to the real world. For reducing the domain shifts, Wang *et al.* [17] are the first to propose a crowd counting via domain adaptation method based on CycleGAN [18], which translates synthetic data to photo-realistic scenes and then apply the trained model in the wild. In this paper, we also focus on Domain-Adaptive Crowd Counting (DACC), which attempts to transfer the useful knowledge for crowd counting from a source domain (the synthetic data) to a target domain (the real world).

However, there are **two problems** in the CycleGAN-style [18], [19], [17], [20] adaptation methods: 1) output some distorted translations and lose many textures and local structured patterns (these features are key characteristics for congested crowd scenes), which produce coarse density map. The left box in Fig. 1 shows the three types of false cases (Red: lost textures, Green: distorted data, and Blue: lost local pattern). 2) mistakenly estimate response values for unseen background objects in the target domain so that the prediction map is very coarse and inaccurate. The right box in Fig. 1 demonstrates some mis-estimations of the background.

For the first problem, the main reason is that CycleGAN only classifies the translated and recalled results at the image level and treats image translation as an entire process. In practice, we find that different domains have common crowd contents, namely person's structure features and crowd distribution patterns, which is regarded as "domain-shared features". Besides, different domains have own unique scene attributes, named as "domain-independent features", which may be caused by different factors such as backgrounds, sensors' setting. Motivated by this discovery, we propose a two-step chain architecture to segregate the two types of features, named as Inter-domain Features Segregation (IFS). It firstly extracts domain-shared features f. Next, by decorating f with the domain-independent features of domain $\mathcal{T}$, IFS reconstructs the like-$\mathcal{T}$ images. For further maintaining the local patterns and texture features, we carefully design multi-scale adversarial translation loss and content-aware loss. Compared with the traditional adversarial loss and Cycle loss, the proposed losses can significantly reduce distortions and retain image contents during the translation.

For the second problem, we present a re-training scheme base on the density reconstruction. In the counting field, the ground-truth of density map is generated by using a

Manuscript received January 12, 2021; revised August 06, 2021; accepted October 22, 2021. This work was supported by the National Natural Science Foundation of China under Grant U1864204, 61773316, 61632018, and 61825603.

J. Gao, T. Han, Y. Yuan and Q. Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China. E-mails: gjy3035@gmail.com, hantao10200@mail.nwpu.edu.cn, crabwq@gmail.com, y.yuan1.ieee@gmail.com. Q. Wang is the corresponding author.

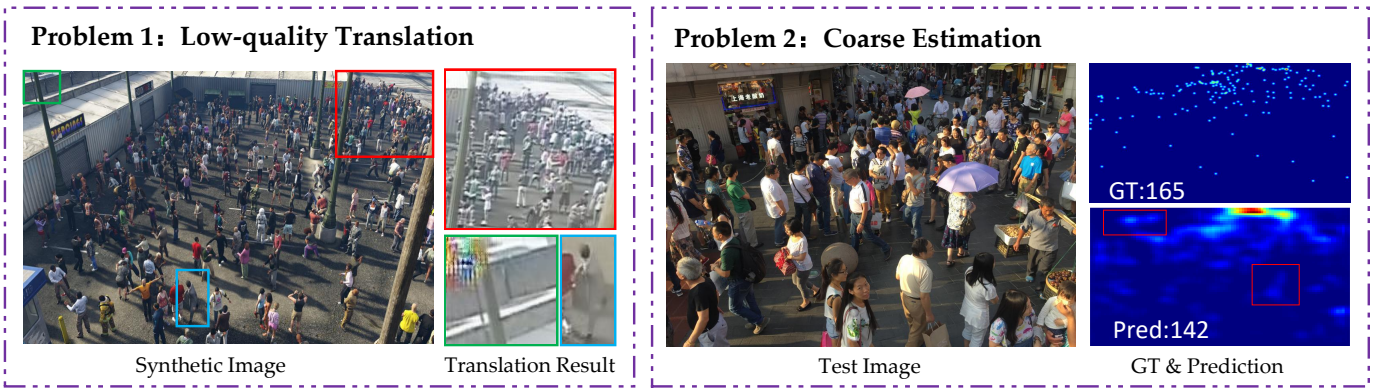


Fig. 1: Existing problems in the current domain-adaptive crowd counting.

Gaussian kernel from the head position. According to this prior, we attempt to find the most likely locations of heads by comparing the similarity between the coarse map and the standard Gaussian kernel. Consequently, pseudo density maps are reconstructed. Then a final counter is trained on the target images and the pseudo maps, which performs better in the real world than the coarse model.

As a summary, the key contributions of this paper are:

1) Propose a two-step image translation to segregate inter-domain features, and design two effective types of losses, which can extract/retain crowd contents and yield high-quality photo-realistic crowd images.
2) Exploit Gaussian prior to reconstruct pseudo labels according to the coarse results. Based on them, retrain a fine counter to further enhance the density quality and counting performance.
3) The proposed method outperforms the state-of-the-art results in the domain-adaptive crowd counting from synthetic data to the real world.

II. Related Works

A. Crowd Counting

Supervised Learning. Early methods for crowd counting focus on extracting hand-crafted features (such as Harr [21], HOG [22], texture features [23], etc.) to regress the number of people [24], [25], [26]. Recently, many object counting researches are based on CNN methods. Some researchers design network structures to enhance multi-scale feature extraction capabilities [27], [28]. Zhang *et al.* [27] propose a multi-column CNN by combining different kernel sizes. Onoro-Rubio and López-Sastre [28] present a multi-scale Hydra CNN, which performs the density prediction in different scenes. Some works [29], [30], [31], [32] exploit contextual information to boost counting performance. Sindagi *et al.* [29] extract global and local feature to aid the density estimation, and Liu *et al.* [30] present a context-aware CNN, designing a multi-stream with different respective fields after a VGG backbone. The rest works [33], [34], [35] fuse multi-stage features to achieve accurate counting. Idrees *et al.* [33] combine the results of different stages to predict the density map and head localization. Jiang *et al.* [34] design a trellis encoder-decoder architecture to incorporate the features from multiple decoding paths. Liu *et al.* [35] present a Structured Feature Enhancement Module (SFEM) using Conditional Random Field (CRF) to refine the features of different stages.

Counting for Scarce Data. In addition to the aforementioned supervised methods, some approaches dedicate to handling the problem of scarce data. Wang *et al.* [17] construct a large-scale synthetic crowd dataset, including more than $15,000$ images, ~ 7.5 million instances. Recently, two real-world crowd datasets are released, namely JHU-CROWD [36] ($4,250$ images, ~ 1.1 million instances) and Crowd Surveillance [37] ($13,945$ images, ~ 0.4 million annotations). By comparing them, the amount of labeled real data is far from that of synthetic data. Besides, collecting and annotating real data is an expensive and difficult assignment. Thus, some researchers remedy this problem from the methodology. Liu *et al.* [38] propose a self-supervised ranking scheme as an auxiliary to improve performance. Sam *et al.* [39] present an almost unsupervised method, of which 99.9% parameters in the proposed auto-encoder is trained without any label. Olmschenk *et al.* [40] enlarge the data by utilizing Generative Adversarial Networks (GAN). To fully escape from manually labeled data and simultaneously attain an accepted result, Wang *et al.* [17] present a crowd counting via domain adaptation method, which is easy to land in practice from the perspectives of performance and costs.

B. Domain-adaptive Vision Tasks

Considering that there are not many works about domain adaptation in crowd counting, thus this section reviews other applications, such as classification, segmentation, etc. Some methods [41], [42], [43] adopt the Maximum Mean Discrepancy [44] to alleviate domain shift in the field of image classification. After some synthetic segmentation datasets [45], [46] are released, a few works [47], [48], [49], [50] adopt adversarial learning to reduce the pixel-wise domain gap. Benefiting from the power of CycleGAN [18], some scholars [19], [20] utilize it to translate synthetic images to realistic data. Recently, some researchers attempt to disentangle the image content and style to translate images [51], [52], [53], [54]. From these works for image classification and segmentation tasks, they only focus on global styles and object-level structures. For counting task, in addition to the style and structures, we also devote to

preserving the consistency of local pattern and textures. Thus, we start from the translation architecture and loss designation to achieve our goal.

Different from the traditional segregation methods [55], [52] that extract a simple domain code to generate images: in order to reconstruct the image details, we uses two different generators that contains a large number of neurons. The richer domain-independent attributes are stored in the generators than the simple domain code.

Different from the previous Cycle-Consistent methods [18], [19], [20] that only constrain the original image and the recalled image by a cycle-consistent loss: to maintain the key content during the translation process, we should regularize the original image and the translated image. To this end, we propose a content-aware consistency loss to guarantee translated data do not lose the original image content, local pattern and texture information.

Different from the previous feature-level adversarial learning algorithms [50], [49] that directly learn the domain-invariant features: the proposed method is based on high-quality image translation, which is more interpretable. Besides, it can be treated as data augmentation. Cai *et al.* [56] propose a two-stage domain adaptation method, which first utilizes adversarial learning in multi-level feature to strengthen target domain's adaptability, and then uses the predicted density maps in the first stage as the pseudo-labels to retrain the counter.

III. Our Method

Here, the proposed DACC is explained from the perspective of data flow. Specifically, a source domain provides crowd images $I_{\mathcal{S}}$ with the labeled density maps $A_{\mathcal{S}}$; and a target domain only provides images $I_{\mathcal{T}}$. The purpose is to get the prediction density maps $\hat{A}_{\mathcal{T}}$ according to given the $I_{\mathcal{S}}$, $A_{\mathcal{S}}$ and $I_{\mathcal{T}}$. To help the reader understand, some of the symbol used behind are concluded in Table I

TABLE I: Some beforehand annotations of involved symbol.

Symbol	Explanation
$\mathcal{S}$	source domain (synthetic data)
$\mathcal{T}$	target domain (real-world data)
$\boldsymbol{G}_c$	domain-shared feature extractor (see Fig. 2)
$\boldsymbol{G}_{to\mathcal{S}}$	source domain decoder (see Fig. 2)
$\boldsymbol{G}_{to\mathcal{T}}$	target domain decoder (see Fig. 2)
$\boldsymbol{D}_c$	domain-shared feature discriminator (see Fig. 2)
$\boldsymbol{D}_{to\mathcal{S}}$	source domain discriminator (see Fig. 2)
$\boldsymbol{D}_{to\mathcal{T}}$	target domain discriminator (see Fig. 2)

A. High-quality Image Translation for Crowd Counting

Image translation aims to translate source images $I_{\mathcal{S}}$ to like-target data $\hat{I}_{\mathcal{S}to\mathcal{T}}$. At the same time, the latter is supposed to contain the key crowd contents of the former. Inspired by the disentangled representation [51], [52], we propose an Inter-domain Features Segregation (IFS) framework to separate the crowd contents and domain-independent attributes. Finally, exploiting the translated images and source labels, we train a coarse crowd counter.

3.1.1 Inter-domain Features Segregation

Assumption. For crowd scenes of different domains, some essential contents are shared, such as the structure information of persons, the arrangement of congested crowds. Meanwhile, each domain has its private attributes, such as different backgrounds, image styles, viewpoints. Thus, we assume that *a source domain shares a latent feature space with any other target domain, and each domain has its independent attribute*.

Model Overview. Based on this assumption, the purpose of IFS is supposed to separate common crowd contents and private attributes without overlapping. It consists of two components, a domain-shared features extractor $\boldsymbol{G}_c$ and two domain-specific decoders $\boldsymbol{G}_{to\mathcal{S}}$ and $\boldsymbol{G}_{to\mathcal{T}}$ for source and target domains. To separate two types of features, we design three corresponding adversarial discriminators for them. The discriminators attempt to distinguish which domain the outputs of $\boldsymbol{G}_c$, $\boldsymbol{G}_{to\mathcal{S}}$ and $\boldsymbol{G}_{to\mathcal{T}}$ come from. By optimizing generators and discriminators in turns, $\boldsymbol{G}_c$ can extract domain-shared features, and $\boldsymbol{G}_{to\mathcal{S}}$, $\boldsymbol{G}_{to\mathcal{T}}$ can reconstruct source domain-like or target domain-like crowd scenes according to the outputs of feature extractor. Consequently, the domain-shared features are extracted explicitly and the domain-specific features are implicitly contained in source domain decoder and target domain decoder.

Domain-shared Features Extractor $\boldsymbol{G}_c$. Based on the above assumption, it is important to ensure that feature extractor extracts similar feature distributions for the samples from different domains (namely $i_{\mathcal{S}} \in I_{\mathcal{S}}$ and $i_{\mathcal{T}} \in I_{\mathcal{T}}$). To this end, we introduce a feature-level adversarial learning for the $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$ produced by the features extractor, of which is corresponding to source domain and target domain, respectively. Specifically, training a discriminator $\boldsymbol{D}_c$ to distinguish whether the features come from source domain or target domain. At the same time, updating the parameters of feature extractor to fool $\boldsymbol{D}_c$ by using the loss of the inverse discrimination result. Consequently, $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$ are very similar and share the same feature space.

Domain-specific Decoders $\boldsymbol{G}_{to\mathcal{S}}$ and $\boldsymbol{G}_{to\mathcal{T}}$. The proposed $\boldsymbol{G}_c$ can extract the features that share the same feature space, but it does not mean that they are key contents mentioned in Assumption. Thus, we propose two domain-specific decoders for domain $\mathcal{S}$ and $\mathcal{T}$, which reconstruct images like own domain according to the outputs of feature extractor. On the one hand, this process encourages feature extractor to extract effective domain-shared features. On the other hand, it makes $\boldsymbol{G}_{to\mathcal{S}}$ and $\boldsymbol{G}_{to\mathcal{T}}$ contain the domain-independent attributes.

To achieve the above goals, we introduce adversarial networks $\boldsymbol{D}_{to\mathcal{S}}$ and $\boldsymbol{D}_{to\mathcal{T}}$ for each domain-specific decoders, respectively. They attempt to determine which domain is the origin of reconstructed images. Taking $\{f_{\mathcal{S}}, f_{\mathcal{T}}, \boldsymbol{G}_{to\mathcal{T}}, \boldsymbol{D}_{to\mathcal{T}}\}$ as an example, feed $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$ into $\boldsymbol{G}_{to\mathcal{T}}$, then attain $\hat{i}_{\mathcal{S}to\mathcal{T}}$ and $\hat{i}_{\mathcal{T}to\mathcal{T}}$ respectively. $\boldsymbol{D}_{to\mathcal{T}}$ aims to distinguish the domains of $\hat{i}_{\mathcal{S}to\mathcal{T}}$ and $\hat{i}_{\mathcal{T}}$. Similar to the above feature-level adversarial training, the loss of the inverse discrimination result is used to update the $\boldsymbol{G}_c$ and $\boldsymbol{G}_{to\mathcal{T}}$. As a result, the photo-realistic

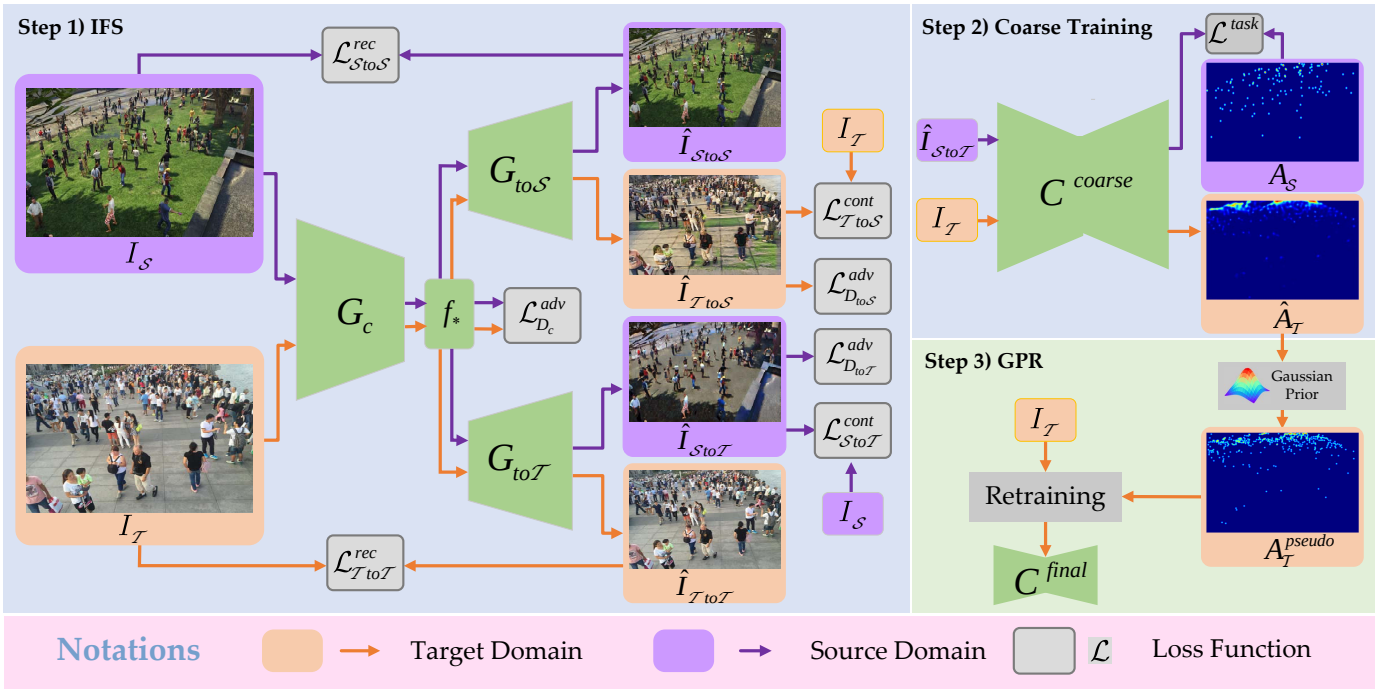


Fig. 2: The flowchart of our proposed method, which consists of three components: 1) IFS translates $I_{\mathcal{S}}$ to $I_{\mathcal{S}to\mathcal{T}}$; 2) Train the coarse counter $\boldsymbol{C}^{coarse}$ using $I_{\mathcal{S}to\mathcal{T}}$ and $A_{\mathcal{S}}$; 3) After $\boldsymbol{C}^{coarse}$ converges via iteratively optimizing Step 1) and 2), reconstruct the pseudo map $A_{\mathcal{T}}^{pseudo}$ from $\boldsymbol{C}^{coarse}$'s predictions $\hat{A}_{\mathcal{T}}$ and retrain the final counter $\boldsymbol{C}^{final}$ using $I_{\mathcal{T}}$ and $A_{\mathcal{T}}^{pseudo}$. Limited by the paper space, the three discriminators are not shown in the figure.

image $\hat{i}_{\mathcal{S}to\mathcal{T}}$ is generated to fool $\boldsymbol{D}_{to\mathcal{T}}$.

3.1.2 Coarse Training for Crowd Counting

After generating the translated images $\hat{I}_{\mathcal{S}to\mathcal{T}}$, the coarse counter $\boldsymbol{C}^{coarse}$ is trained on $\hat{I}_{\mathcal{S}to\mathcal{T}}$ and $A_{\mathcal{S}}$ by using the traditional supervising regression method. In practice, given a batch of translation results in each iteration of IFS, $\boldsymbol{C}^{coarse}$ will be trained once. In other word, the image translation model and the coarse counter are trained together.

3.1.3 Loss Functions

To train the proposed framework, in each iteration, the discriminators $\boldsymbol{D}_c$, $\boldsymbol{D}_{to\mathcal{S}}$ and $\boldsymbol{D}_{to\mathcal{T}}$ are updated using an adversarial loss; then update the parameters of $\boldsymbol{G}_c$, $\boldsymbol{G}_{to\mathcal{S}}$, $\boldsymbol{G}_{to\mathcal{T}}$, and $\boldsymbol{C}^{coarse}$ by optimizing following functions:

$$\mathcal{L} = \mathcal{L}^{task} + \alpha\mathcal{L}^{adv}_{\boldsymbol{D}_c} + \beta\mathcal{L}^{adv}_{\boldsymbol{D}_{to\mathcal{S}}} + \gamma\mathcal{L}^{adv}_{\boldsymbol{D}_{to\mathcal{T}}} + \mathcal{L}^{cons}, \quad (1)$$

where the first item is task loss for counting, the middle threes are adversarial loss for three discriminators, and the last item is the consistency loss. By repeating the above training, the models will be obtained. Next, we will explain the concrete definitions of them. Note that θ_* means that the parameters of the model $*$.

Task Loss For the counting task, we train $\boldsymbol{C}^{coarse}$ via optimizing $\mathcal{L}^{task}(\theta_{\boldsymbol{C}^{coarse}})$, a standard MSE loss.

Feature-level Adversarial Loss To effectively extract domain-shared features, we minimize feature-level LSGAN loss [57] to train $\boldsymbol{D}_c$. The loss and inverse loss is denoted by $\mathcal{L}_{\boldsymbol{D}_c}$ and $\mathcal{L}^{adv}_{\boldsymbol{D}_c}$, respectively.

Multi-scale Translation Adversarial Loss We find that the traditional methods are prone to generating weird data that contain distorted color distribution. The main reason is that the adversarial training is unstable and it causes that some neurons are sensitive to specific data. To alleviate this problem, we propose a multi-scale translation adversarial loss (MS Ad for short). which adopts a full convolution discriminator to distinguishes the domains of two images under the different image scales. It is a MSE-like loss and has been proved in LSGAN [57] with better stability compared with the Cross-entropy loss. Take $\boldsymbol{D}_{to\mathcal{S}}$ as an example, $\mathcal{L}_{\boldsymbol{D}_{to\mathcal{S}}}$ and $\mathcal{L}^{adv}_{\boldsymbol{D}_{to\mathcal{S}}}$ are formulated as:

$$\mathcal{L}_{\boldsymbol{D}_{to\mathcal{S}}}(\theta_{\boldsymbol{D}_{to\mathcal{S}}}) = \frac{1}{2}\sum_{l=1}^{2}\Big\{||\boldsymbol{D}_{to\mathcal{S}}(i^l_{\mathcal{S}}) - 0||^2 + ||\boldsymbol{D}_{to\mathcal{S}}(i^l_{\mathcal{T}to\mathcal{S}}) - 1||^2\Big\}, \quad (2)$$

and

$$\mathcal{L}^{adv}_{\boldsymbol{D}_{to\mathcal{S}}}(\theta_{\boldsymbol{G}_c}, \theta_{\boldsymbol{G}_{to\mathcal{S}}}) = \frac{1}{2}\sum_{l=1}^{2}||\boldsymbol{D}_{to\mathcal{S}}(i^l_{\mathcal{T}to\mathcal{S}}) - 0||^2, \quad (3)$$

where $l = 1, 2$ respectively represents the size of inputs, namely 0.5x and 1.0x. During the training, $\boldsymbol{D}_{to\mathcal{S}}$ attempts to distinguish the origins of $i_{\mathcal{S}}$ and $i_{\mathcal{T}to\mathcal{S}}$. At the same time, by optimizing $\mathcal{L}^{adv}_{\boldsymbol{D}_{to\mathcal{S}}}$, the $\boldsymbol{G}_c$ and $\boldsymbol{G}_{to\mathcal{S}}$ are updated to generate like-target images that can confuse $\boldsymbol{D}_{to\mathcal{S}}$. Similarly, there are $\mathcal{L}_{\boldsymbol{D}_{to\mathcal{T}}}(\theta_{\boldsymbol{D}_{to\mathcal{T}}})$ and $\mathcal{L}^{adv}_{\boldsymbol{D}_{to\mathcal{T}}}(\theta_{\boldsymbol{G}_c}, \theta_{\boldsymbol{G}_{to\mathcal{T}}})$ to train $\boldsymbol{D}_{to\mathcal{S}}$ and $\{\boldsymbol{G}_c, \boldsymbol{G}_{to\mathcal{S}}\}$, respectively.

Consistency Loss Mainstream translation methods have two data flows: recall process ($i_{\mathcal{S}} \rightarrow \hat{i}_{\mathcal{S}to\mathcal{S}}$) and the translation

process ($i_{\mathcal{T}} \rightarrow \hat{i}_{\mathcal{T}to\mathcal{S}}$). For the former, the researchers [18] usually regularize the data using pixel-wise consistency loss (namely L2 Loss). However, for image translation task, the ultimate goal is translating images instead of recalling images. Many works ignore the latter so that the model loses the original content and detailed features.

To remedy this problem, we attempt to design a loss function to constrain high-level image content. The L2 Loss in recall process is improper, because it only measure the pixel-wise distance, which is anti-translation operation. Therefore, we propose a content-aware consistency loss to regularize $i_{\mathcal{T}}$ and $\hat{i}_{\mathcal{T}to\mathcal{S}}$. To be specific, we adopt perceptual losses [58] to formulate the difference of feature maps extracted by a pre-trained classification model VGG-16 [59], which are named as $\mathcal{L}^{cont}_{\mathcal{T}to\mathcal{S}}(\theta_{\boldsymbol{G}_c}, \theta_{\boldsymbol{G}_{to\mathcal{S}}})$ and $\mathcal{L}^{cont}_{\mathcal{S}to\mathcal{S}}(\theta_{\boldsymbol{G}_c}, \theta_{\boldsymbol{G}_{to\mathcal{S}}})$. It effectively maintains low-level local features and high-level crowd contents of the original image. Similarly, there are $\mathcal{L}^{rec}_{\mathcal{T}to\mathcal{T}}(\theta_{\boldsymbol{G}_c}, \theta_{\boldsymbol{G}_{to\mathcal{T}}})$ and $\mathcal{L}^{cont}_{\mathcal{S}to\mathcal{T}}(\theta_{\boldsymbol{G}_c}, \theta_{\boldsymbol{G}_{to\mathcal{T}}})$ to regularize the outputs of $\boldsymbol{G}_{to\mathcal{T}}$.

Finally, $\mathcal{L}^{cons}$ in Eq. 1 is the sum of the above four consistency losses.

B. Gaussian-prior Reconstruction

In the field of crowd counting, the ground-truth of density map is generated using head locations and Gaussian kernel [25]. The goal of Gaussian-prior Reconstruction (GPR) is to find the most likely head locations via comparing the coarse map and the standard kernel. After this, the pseudo map is reconstructed and used to train a final counter on the target domain.

Density Map Generation Firstly, we briefly review the generation process of density maps in traditional supervised methods. In the field of counting, the original label form is a set of heads positions $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \ldots, (x_N, y_N)\}$. Take a sample (x_i, y_i) as an example, it is treated as a delta function $\delta(x - x_i, y - y_i)$. Therefore, the position set can be formulated as:

$$H(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{N} \delta(x - x_i, y - y_i). \tag{4}$$

For getting the density map, we convolve $H(\boldsymbol{x}, \boldsymbol{y})$ with a Gaussian function $G_{k,\sigma}$, where k is the kernel size and σ is the standard deviation. In practice, $G_{k,\sigma}$ is regarded as a discrete Gaussian Window $W_{k,\sigma}$ with the size of $k \times k$. To be specific, the value of position (u, v) in $W_{k,\sigma}$ is defined as $w(u, v) = e^{-D^2(u,v)/2\sigma^2}$, where $D(u, v)$ is the distance from (u, v) to the window center. It is defined as:

$$D(u, v) = \left[((u-(k+1)/2)^2 + (v-(k+1)/2)^2\right]^{1/2}. \tag{5}$$

In the experiments, we set k as 15 and σ as 4.

Density Map Reconstruction A standard map is recalled according to the coarse result $\hat{a}_{\mathcal{T}}$. It consists of three steps: 1) compute probability map at the pixel level, of which each pixel represents its confidence as a Gaussian kernel's center; 2) iteratively select a maximum-probability candidate point and update the probability map in turns; 3) generate pseudo labels based on candidate points.

Here, we detailedly explain the generation of the probability map. Take a pixel (x_i, y_i) in $\hat{A}_{\mathcal{T}}$ as the center, cropping a window $\hat{A}^{(x_i,y_i)}_{\mathcal{T}}$ with the size of $k \times k$. Then measuring the similarity between $\hat{A}^{(x_i,y_i)}_{\mathcal{T}}$ and $W_{k,\sigma}$ using following formulation:

$$P(x_i, y_i) = \frac{1}{1 + \|\hat{A}^{(x_i,y_i)}_{\mathcal{T}} - W_{k,\sigma}\|_1}, \tag{6}$$

where $W_{k,\sigma}$ is a discrete Gaussian Window with the size of $k \times k$, $P(x_i, y_i) \in [0, 1]$ and the higher value means that it is closer to the $W_{k,\sigma}$. Finally, the probability map P is obtained. The generation flow is shown in Fig.3, and the computation process is demonstrated in Algorithm 1.

Algorithm 1 Algorithm for generating pseudo labels.

Require: Coarse map $\hat{A}_{\mathcal{T}}$, Gaussian Window $W_{k,\sigma}$.
Ensure: Pseudo label map $A^{pseudo}_{\mathcal{T}}$.

```
Count the number of people, N̂ = int(sum(Â_T));
Compute the probability map P for Â_T^coarse with Eq. 6;
for j = 1 to N̂ do
    Get a candidate point (x̂_j, ŷ_j) = arg max_{(x̂_j, ŷ_j)} (P(x̂_j, ŷ_j));
    Crop a window Â_T^{(x̂_j, ŷ_j)} with the center (x̂_j, ŷ_j) from Â_T;
    Update Â_T^{(x̂_j, ŷ_j)} = Â_T^{(x̂_j, ŷ_j)} − W_{k,σ};
    Place Â_T^{(x̂_j, ŷ_j)} back to Â_T;
    Recompute P's region where changes occur in Â_T;
end for
Generate the map A_T^pseudo with {(x̂_1, ŷ_1), ..., (x̂_N̂, ŷ_N̂)}.
return A_T^pseudo.
```

Re-training Scheme Although the above reconstruction can effectively prompt the density quality, it may generate a few mistaken head labels from the coarse map. In addition, its time complexity is $O(n)$, which is not efficient. To remedy these problems, we re-train a final counter $\boldsymbol{C}^{final}$ using $I_{\mathcal{T}}$ and $A^{pseudo}_{\mathcal{T}}$ based on the $\theta_{\boldsymbol{C}^{coarse}}$. The error labels will be alleviated as the model converges. During the test phase, the $\boldsymbol{C}^{final}$ is performed to directly more high-quality predictions than the coarse results.

C. Network Architecture

This section briefly describes our network architectures. $\boldsymbol{G}_c$ consists of four residual blocks and outputs 512-channel feature map with the 1/4 size of inputs. $\boldsymbol{G}_{to\mathcal{S}}$ and $\boldsymbol{G}_{to\mathcal{T}}$ have the same architecture, including six convolutional/de-convolutional layers. For the discriminators, they are all designed as a five-layer convolution network. The counters utilize the first 10 layers of VGG-16 [59], and up-sample to the original size via a series of de-convolutional layers. All detailed configurations of the networks are shown in supplementary materials, and the code will be released as soon as possible.

D. Implementation details

Parameter Setting During the training process of IFS, the weight parameters α, β, and γ in Eq.1 are set to 0.01, 0.1, and

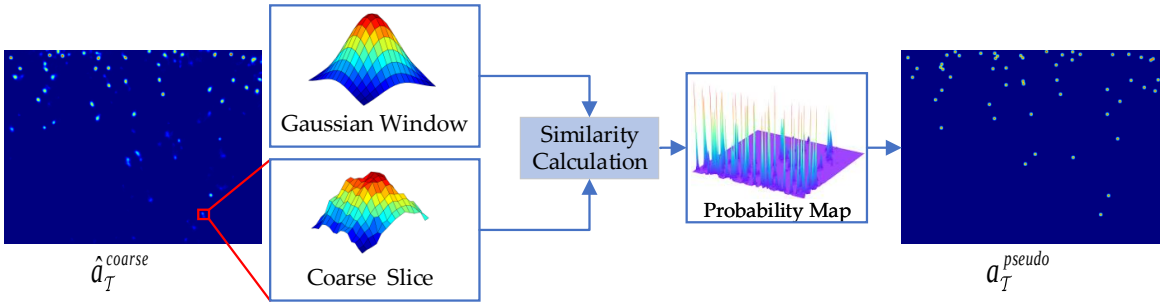


Fig. 3: The generation process of pseudo labels.

0.1, respectively. Due to the limited memory, in each iteration, we input 4 source images and 4 target images with a crop size of 480×480. Adam algorithm [60] is performed to optimize the networks. The learning rate for the IFS models is set as 10^{-4}, and the learning rate for $\boldsymbol{C}^{coarse}$ is initialized as 10^{-5}. After $4,000$ iterations, we stop updating the IFS models, but continue to update the $\boldsymbol{C}^{coarse}$ until it converges. For GPR process, $\boldsymbol{C}^{final}$'s learning rate is set as 10^{-5}. Our code is developed based on the C^3 Framework [61] on NVIDIA GTX 1080Ti GPU.

Scene Regularization In other fields of domain adaptation, such as semantic segmentation, the object distribution in street scenes is highly consistent. Unlike this, current crowd real-world datasets are very different in terms of density. For avoiding negative adaptation, we adopt a scene regularization strategy proposed by [17]. In other word, we manually select some proper synthetic scenes from GCC as the source domain for different target domains. Due to no experiment on UCSD [62] and Mall[63] in SE CycleGAN [17], we define the scene regularization for them. The detailed information is shown in the supplementary.

IV. Experimental Results

A. Evaluation Criteria

Following the convention, we utilize Mean Absolute Error (MAE) and Mean Squared Error (MSE) to measure the counting performance of models, which are defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|^2}, \quad (7)$$

where N is the number of images, y_i is the groundtruth number of people and $\hat{y}_i$ is the estimated value for the i-th image. Besides, PSNR and SSIM [64] is adopted to evaluate the quality of density maps.

B. Datasets

For verifying the proposed domain-adaptive method, the experiments are conducted from GCC [17] to another six real-world, namely Shanghai Tech Part A/B [27], UCF-QNRF [33], WorldExpo'10 [65], Mall [62] and UCSD [63].

GCC is a large-scale synthetic dataset, which consists of still $15,212$ images with a resolution of 1080×1920.

Shanghai Tech Part A is a congested crowd dataset, of which images are from a photo-sharing website. It consists of 482 images with different resolutions.

Shanghai Tech Part B is captured from the surveillance camera on the Nanjing Road in Shanghai, China. It contains 716 samples with a resolution of 768×1024.

UCF-QNRF is an extremely congested crowd dataset, including 1,535 images collected from Internet, and annotating in 1,251,642 instances.

WorldExpo'10 is collected from 108 surveillance cameras in Shanghai 2010 WorldExpo, which contains $3,980$ images with a size of 576×720.

Mall is collected using a surveillance camera installed in a shopping mall, which records the $2,000$ sequential frames with a resolution of 480×640.

UCSD is an outdoor single-scene dataset collected from a video camera at a pedestrian walkway, which contains $2,000$ image sequences with a size of 158×238.

C. Module-level Ablation Study on Shanghai Tech Part A

We conduct a group of detailed ablation study to verify the effectiveness of our proposed models on Shanghai Tech Part A. To be specific, the different models' configurations are explained as follows:

Table II reports the quantitative results of different module fusion methods.

NoAdpt: Train the counter on the original GCC.

IFS-a: Train the translated GCC of IFS w/o feature-level adversarial learning.

IFS-b: Train the translated GCC of IFS with feature-level adversarial learning.

IFS-b + GPR-a: Reconstruct pseudo labels using the results of the counter in IFS-b.

IFS-b + GPR-b: Retrain the counter with the pseudo labels of IFS-b + GPR-a. It is the full model of this paper, namely the proposed DACC.

Analysis of IFS From the table, the methods with adaptation far exceed NoAdpt, which shows the effectiveness of domain adaptation. By comparing the results IFS-a and IFS-b, the errors are significantly reduced (MAE/MSE: from 127.3/190.6 to 120.8/184.6). It indicates that the feature-level adversarial learning effectively facilitates the segregation of inter-domain features.

Analysis of GPR When introducing GPR-a into IFS-b, the counting errors are slightly different (MAE/MSE: 120.8/184.6

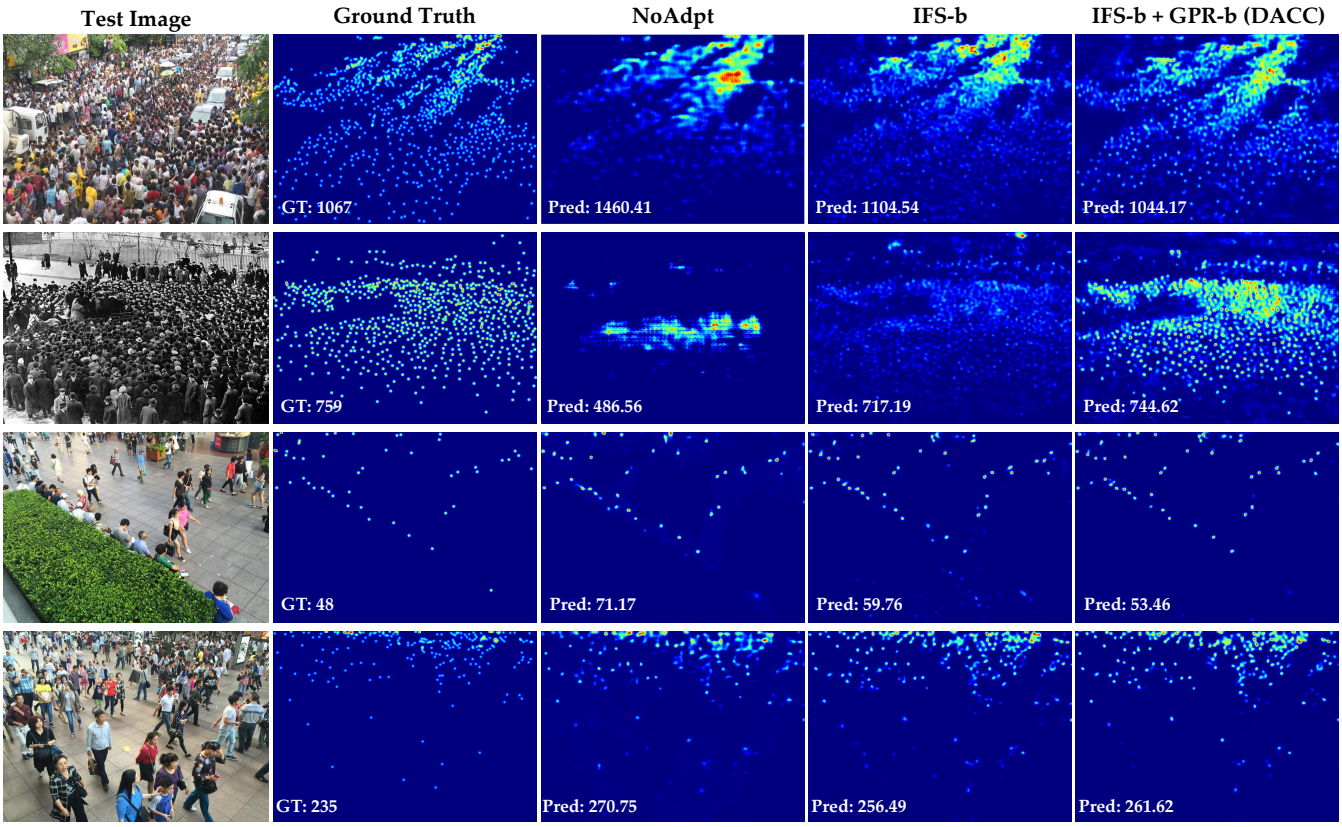


Fig. 4: Exemplar results of adaptation from GCC to Shanghai Tech Part A and B dataset. In the density map. "GT" and "Pred" represent the number of ground truth and prediction, respectively. Row 1 and 2 come from ShanghaiTech Part A, and others are from Part B.

TABLE III: The performance of the proposed different loss combinations on UCF-QNRF.

Loss Combinations	CycleGAN Structure				IFS-b Structure			
	MAE	MSE	PSNR	SSIM	MAE	MSE	PSNR	SSIM
Ad + C (original)	257.3	400.6	20.80	0.480	243.9	392.6	20.77	0.607
MS Ad + C	232.9	394.0	20.97	0.575	221.8	385.9	21.23	0.642
ad + c + SE	230.4	384.5	21.03	**0.660**	225.3	281.7	21.57	**0.690**
ad + c + CA	223.7	381.7	21.09	0.612	215.8	361.0	21.77	0.676
MS Ad + C + CA	**218.1**	**380.0**	**21.17**	0.624	**211.7**	**357.9**	**21.94**	0.687

TABLE II: The performance of the proposed different models on Shanghai Tech Part A.

Method	Shanghai Tech Part A			
	MAE	MSE	PSNR	SSIM
NoAdpt	206.7	297.1	18.64	0.335
IFS-a	127.3	190.6	21.80	0.458
IFS-b	120.8	184.6	21.41	0.466
IFS-b + GPR-a	120.6	184.4	19.73	**0.760**
IFS-b + GPR-b (DACC)	**112.4**	**176.9**	**21.94**	0.502

v.s. 120.6/184.4). The main reason is that the rounding operation for counting number in Line 1 of Algorithm 1. It is a double-edged sword, which maybe decrease or increase the errors. The slight performance fluctuations are not important. Our concern is to improve the quality of density map and remove some misestimations by the further retraining scheme. Correspondingly, since GPR-a generates the standard pseudo labels, SSIM achieves the value of 0.760, far more than the previous result of 0.466. After retraining a final counter, the mistaken estimations in the background are dramatically suppressed. As a result, the MAE and MSE of DACC are further reduced (MAE/MSE: 120.6/184.4 v.s. 112.4/176.9).

Visualization Results Fig. 4 shows the visualization results of the proposed step-wise models (NoAdpt, IFS-b and IFS-b + GPR-b) on Shanghai Tech Part A and B. From the results of Column 3, NoAdpt only reflects the trend of density distribution. For the second sample, NoAdpt produce a weird density map, which seems to be not consistent with the original image. The main reason is that GCC data are RGB images, but the second sample is a gray-scale scene. The NoAdpt counter fully over-fits the RGB data so that it performs poorly on

gray-scale images. After introducing IFS, the visual results can show the coarse density distribution. For some sparse crowd regions (such as Row 3), the counter yields the fine density map close to the ground truth. Further, the final results of DACC present two advantages in visual perception. Firstly, DACC outputs the more precise density maps, of which points are similar to the standard Gaussian kernel. It will prompt the performance of person localization. Secondly, the mistaken estimations are effectively reduced, especially in Row 3 and 4. In general, DACC's predictions are better than those of other models' in terms of the quantitative and qualitative comparisons.

D. Loss-level Ablation Study on UCF-QNRF

In Section I, we mentioned that our losses can significantly maintain the local patterns and texture features, especially for dense crowd. Here, we will verify our opinion by using the experiments that use different translation models (CycleGAN and our IFS-b) on the most congested dataset: UCF-QNRF. The results are reported in Table III. Here, we describe the notations in the table: "Ad" means the traditional adversarial loss used by CycleGAN, and "C" indicates the standard consistency loss used by CycleGAN [18]. SE is SSIM Embedding loss proposed by SE CycleGAN [17]. "MS Ad" and "CA" are our designed multi-scale adversarial loss and context-aware consistency loss.

By comparing the results of the two translation structures, we find the proposed IFS-b is better than CycleGAN under the same training loss combination. The former can extract more effective domain-invariant features than the latter. It evidences that two-stage translation via segregating domain-shared and domain independent features can generate more similar data to real scenes. In Section V-A and Fig. 5, we discuss the translation quality of these the two structures.

Ad Loss *v.s.* MS Ad Loss Different from the previous methods, we attempt to regularize translated images at two resolutions, which facilitates the generator's neurons more robust and remedy the distortion translation outputs. From the final counting errors (MAE), the MS Ad loss has 9.5% (Cycle GAN structure) and 9.1% (IFS-b structure) improvements than the traditional Ad Loss.

CA Loss *v.s.* C Loss C Loss only aims at the recall quality, which does not directly affect the translation. After introducing the CA Loss, the translation quality is prompted and the model attains a better counting performance (13.1% and 11.5% improvements of MAE on Cycle GAN and IFS-b structure).

SE Loss *v.s.* CA Loss Both focus on improving the translation quality on congested regions. By the comparison of the reported results, we find CA loss is superior to SE loss in terms of counting performance. From the perspective of translation quality, SE loss has more significant effect than CA loss (0.660 *v.s.* 0.612). The main reason is that SE loss mainly focuses on local structural similarity instead of high-level image contents.

E. Comparison with the SOTAs on Real-world Datasets

In this section, we perform the experiments of DACC on six mainstream real-world datasets and compare the performance with other domain-adaptive counting methods, such as CycleGAN [18], SE CycleGAN [17], FSC [49], FA [50] and LIDK [56]. Table IV lists the concrete four metrics (MAE↓/MSE↓/PSNR↑/SSIM↑). From it, the proposed DACC outperforms the other methods on all datasets. Take MAE as an example, DACC achieves 112.4, 13.1, 203.5, 17.4,1.76 and 2.31 on the six real-world datasets. In terms of some results on density quality, FSC is better than ours. The main reason is that FSC uses the crowd mask to align the semantic consistency. The extra label effectively reduce the estimation error in the background regions. More visualization results on other datasets are shown in the supplementary.

Compared with NoAdpt, DACC significantly reduces the counting errors. Take MAE as an example, DACC reduce the estimation errors by 45.7%, 47.2%, and 30.5% on the first three dense datasets, respectively. On the sparse WorldExpo'10, UCSD and Mall datasets, DACC achieves more than 50% improvement, 54.2%, 88.2%, and 61.0% respectively.

For the results on UCSD, we find the PSNR and SSIM of density map is not good. The main reason is that the label definitions of them are different. The source domain (GCC) annotate the head position but the target domain (UCSD) annotate the center position. Nevertheless, it does not affect the evaluation for counting the number of people.

V. Discussions

A. Analysis of Translated Image Quality with SOTAs

This section compares the translation results by visualization and image quality. Fig. 5 demonstrates the three results of CycleGAN, SE CycleGAN and our IFS-b. For the first two methods, they lose the key content and some detailed information, especially in the region red boxes. In addition, they also yield some distorted region in yellow boxes. In general, IFS-b maintains the crowd content well.

TABLE V: The quality comparison of the translated images.

Methods	Mean ↑	Std ↓
NoAdpt	5.467	1.757
CycleGAN	4.935	1.848
SE CycleGAN	5.041	1.846
DACC(ours)	**5.244**	**1.810**

TABLE VI: Performance of the exchange experiments (MAE/MSE).

Data flow: GCC→Mall	Data flow: UCSD→Mall
EXP1: 2.31/2.96	EXP2: 2.85/3.55
EXP1': 2.21/2.85	EXP2': 2.97/3.76

As we all know, evaluating the translation closeness to the target domain is difficult because there is no reference image. Thus, we only assess the translation data from the

TABLE IV: The performance of no adaptation (No Adpt), CycleGAN, SE CycleGAN, FSC, FA and the proposed methods on the six real-world datasets.

Method	DA	Shanghai Tech Part A				Shanghai Tech Part B				UCF-QNRF			
		MAE	MSE	PSNR	SSIM	MAE	MSE	PSNR	SSIM	MAE	MSE	PSNR	SSIM
CycleGAN [18]	✔	143.3	204.3	19.27	0.379	25.4	39.7	24.60	0.763	257.3	400.6	20.80	0.480
SE CycleGAN [17]	✔	123.4	193.4	18.61	0.407	19.9	28.3	24.78	0.765	230.4	384.5	21.03	0.660
SE Cycle GAN (JT) [66]	✔	119.6	189.1	18.69	0.429	16.4	25.8	26.17	0.786	225.9	385.7	21.10	0.642
FSC [49]	✔	129.3	187.6	21.58	**0.513**	16.9	24.7	26.20	0.818	221.2	390.2	**23.10**	0.708
FA [50]	✔	144.6	200.6	-	-	16.0	24.7	-	-	269.5	407.9	-	-
LIDK [56]	✔	-	-	-	-	14.3	22.8	-	-	224.3	375.8	-	-
NoAdpt (ours)	✗	206.7	297.1	18.64	0.335	24.8	34.7	25.02	0.722	292.6	450.7	20.83	0.565
DACC (ours)	✔	**112.4**	**176.9**	**21.94**	0.502	**13.1**	**19.4**	**28.03**	**0.888**	**203.5**	**343.0**	21.99	**0.717**

Method	DA	WorldExpo'10 (only MAE)						UCSD				Mall			
		S1	S2	S3	S4	S5	Avg.	MAE	MSE	PSNR	SSIM	MAE	MSE	PSNR	SSIM
CycleGAN [18]	✔	4.4	69.6	49.9	29.2	9.0	32.4	-	-	-	-	-	-	-	-
SE CycleGAN [17]	✔	**4.3**	59.1	43.7	17.0	7.6	26.3	-	-	-	-	-	-	-	-
SE Cycle GAN (JT) [66]	✔	**4.2**	**49.6**	**41.3**	19.8	**7.2**	**24.4**	-	-	-	-	-	-	-	-
FA [50]	✔	5.7	59.9	19.7	**14.5**	8.1	21.6	2.00	2.43	-	-	2.47	3.25	-	-
NoAdpt (ours)	✗	11.0	49.2	72.2	40.2	17.2	38.0	14.95	15.31	23.66	0.909	5.92	6.70	25.02	0.886
DACC (ours)	✔	4.5	**33.6**	**14.1**	30.4	**4.4**	**17.4**	**1.76**	**2.09**	**24.42**	**0.950**	**2.31**	**2.96**	**25.54**	**0.933**

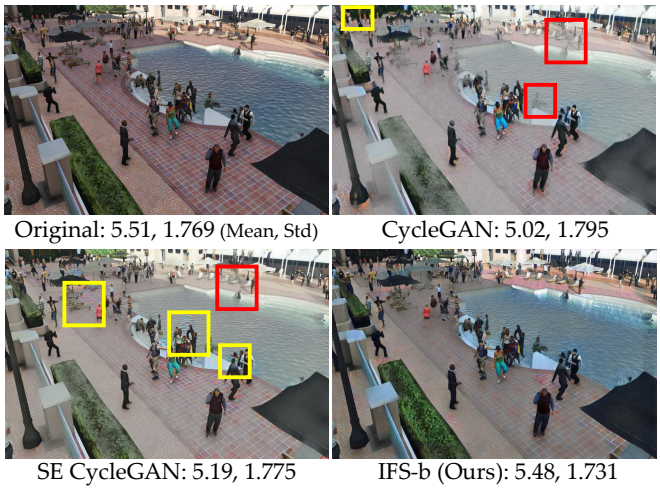


Fig. 5: Comparisons of the adaptation on GCC→ShanghaiTech Part B.

perspective of image quality. Specifically, we utilize a Neural Image Assessment (NIMA) [67], which rates images with a mean score and a standard deviation ("std" for short). Table V reports these two metrics of CycleGAN, SE CycleGAN and the proposed IFS-b on GCC→ShanghaiTech B. We find DACC is better than other translation methods. We also show the NoAdpt results, of which images are the original synthetic GCC data. From the scores of single image in Fig. 5, IFS-b also outperforms CycleGAN-style methods.

B. Visual Analysis of Task Performance with SOTA

To vividly show the effectiveness of DACC, we compare the visual results with the SE CycleGAN [17]. It is emphasized that SE CycleGAN [17] provides visualizations of its counting results on Shanghai Tech Part A and Shanghai Tech Part B. Soo it can be gained directly and compared with the proposed DACC's counting results. Fig. 6 shows the task performance on Shanghai Tech Part A dataset with two methods, which shows the density maps predicted by DACC are highly similar to the ground truth while SE CycleGAN outputs very coarse density maps. This comparison illustrates that the proposed DACC is better than SE CycleGAN in both quality and accuracy.

C. Effectiveness of IFS

In Section III-A, it is mentioned that IFS can effectively separate domain-shared and domain-independent features. Here, we evidence this thought by two groups of exchange experiments. To be specific, select two adaptations with the same target domain, then fix the data and exchange IFS models to translate images. Take two experiments as the examples, 1)EXP1: GCC→Mall and 2)EXP2: UCSD→Mall. We hope to translate GCC data in EXP1 to like-Mall images using the IFS models of EXP2. Then getting the final counter by the translated images and GPR. Finally, the evaluation is conducted on the target data, namely Mall. The above experiment is defined as EXP1'. And vice versa, the other exchange way is named as EXP2'.

The counting results are listed in Table VI and, the translation exemplars are shown in Fig. 7. From them, we find that: given source and target data, exchanging IFS models barely affects the performance of crowd counting and image translation.

D. The performance of counter C via Supervised Learning

In our work, the core is not to design a crowd counter, so we do not pay much attention to the supervised performance

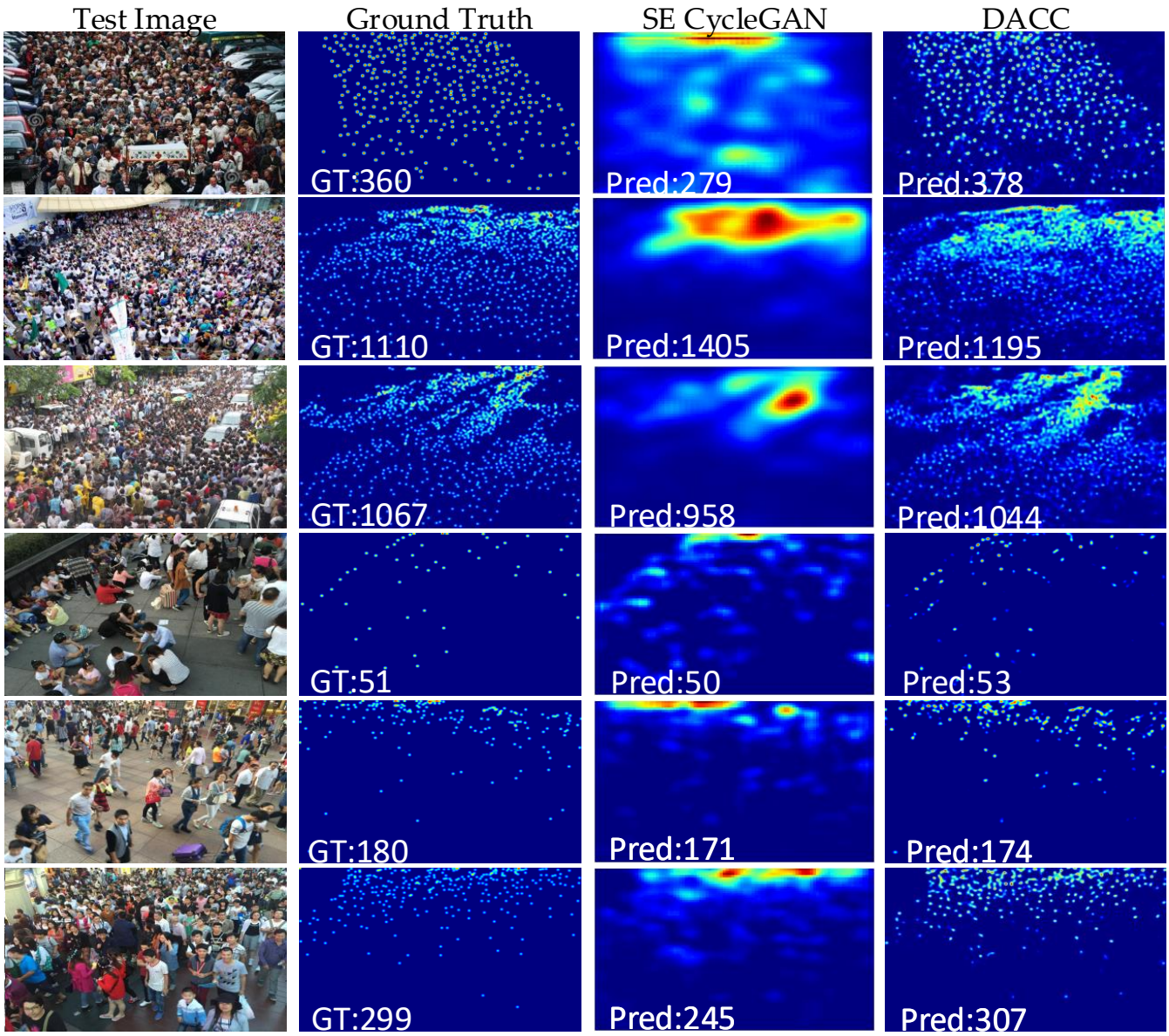


Fig. 6: Exemplar results of adaptation from GCC to Shanghai Tech Part A and B dataset. In the density map. "GT" and "Pred" represent the number of ground truth and prediction, respectively. Row 1 and 2 come from ShanghaiTech Part A, and others are from Part B.

Fig. 7: Visual comparison of the model exchange experiments.

in the target domain. However, in order to prove that the IFS image translation proposed in this paper can effectively reduce the domain gap, we conduct supervised training on several target domains. Table VII compares the performance of counter $\boldsymbol{C}$ between the supervised training in the target domain and domain adaptation. As shown in Table VII, the MAE and MSE of the counter used in this paper are 69.6 and 125.9 respectively on Shanghai Tech Part A, and the MAE and MSE of supervised training on Shanghai Tech Part B are 8.1 and 14.1, respectively. The results in the table show that there is a large gap between no domain adaptation and supervised training, which is significantly reduced after domain adaptation.

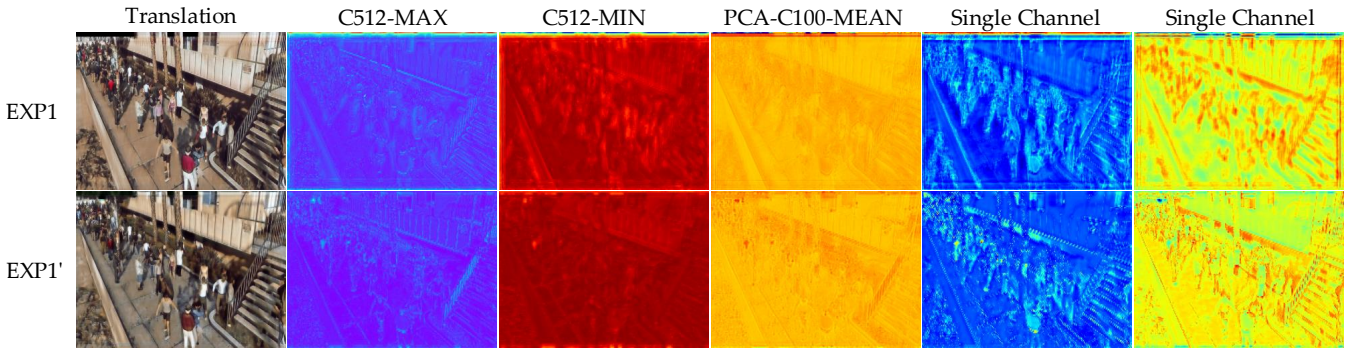


Fig. 8: The feature visualization of $\boldsymbol{G}_c$ in EXP1 and EXP1' with the same source image.

TABLE VII: The comparison of $\boldsymbol{C}$ in the proposed adaptation method and supervised training.

Methods	DA	T-GT	SHT A		SHT B	
			MAE	MSE	MAE	MSE
NoAdapt	✗	✗	206.7	297.1	24.8	34.7
DACC(ours)	✔	✗	112.4	176.9	13.1	19.4
Supervised	✗	✔	**69.6**	**125.9**	**8.1**	**14.1**

E. IFS-b Domain-shared Feature Visualization

In order to verify the effectiveness of our proposed IFS, we conduct a group of exchange experiments in Section 4.5. Here, we show the visualization results at the feature level. To be specific, the domain-shared features of $\boldsymbol{G}_c$ in EXP1 and EXP1' are illustrated in Fig. 8. The first column denotes the image translation results, and the second and third columns respectively represent the maximum, minimum values of each pixel in 512 channels. The fourth column is the average value of each pixel after reducing the original features to 100 channels via PCA. The last two are some similar features selected from 512-channel feature maps. From these visualization results, we find that different $\boldsymbol{G}_c$ from EXP1 and EXP1' can extract similar features for the same image. From Column 5 and 6, there are high responses for the crowd region. In a word, these results evidence that the proposed IFS can extract domain-shared crowd contents.

VI. Conclusion

In this paper, we present a Domain-Adaptive Crowd Counting (DACC) approach without any manual label. Firstly, DACC translates synthetic data to high-quality photo-realistic images by the proposed Inter-domain Features Segregation (IFS). At the same time, we train a coarse counter on translated images. Then, Gaussian-prior Reconstruction (GPR) generate the pseudo labels according to the coarse results. By the re-training scheme, a final counter is obtained, which further refines the quality of density maps on real data. Experimental results demonstrate that the proposed DACC outperforms other state-of-the-art methods for the same task. In future work, we plan to extend IFS on multiple domains so that it can extract more effective and robust crowd contents to improve the counting performance.

References

[1] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[2] X. Li, M. Chen, F. Nie, and Q. Wang, "A multiview-based parameter free framework for group detection," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[3] ——, "Locality adaptive discriminant analysis." in *IJCAI*, 2017, pp. 2201–2207.

[4] K. Kang and X. Wang, "Fully convolutional neural networks for crowd segmentation," *arXiv preprint arXiv:1411.4464*, 2014.

[5] B. Zhou, X. Tang, and X. Wang, "Measuring crowd collectiveness," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3049–3056.

[6] K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, and S. W. Baik, "Efficient deep cnn-based fire detection and localization in video surveillance applications," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 7, pp. 1419–1434, 2018.

[7] B. Zhao, X. Li, and X. Lu, "Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7405–7414.

[8] ——, "Property-constrained dual learning for video summarization," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 10, pp. 3989–4000, 2020.

[9] 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.

[10] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 833–841.

[11] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4031–4039.

[12] 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.

[13] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "Decidenet: Counting varying density crowds through attention guided detection and density estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5197–5206.

[14] 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.

[15] 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.

[16] 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.

[17] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8198–8207.

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

[19] 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.

[20] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. 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.

[21] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005.

[23] G. J. Brostow and R. Cipolla, "Unsupervised bayesian detection of independent motion in crowds," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2006, pp. 594–601.

[24] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Crowd counting using multiple local features," in *2009 Digital Image Computing: Techniques and Applications*, 2009, pp. 81–88.

[25] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Advances in neural information processing systems*, 2010, pp. 1324–1332.

[26] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2547–2554.

[27] 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.

[28] 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.

[29] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1861–1870.

[30] 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.

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

[32] 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*, vol. 30, no. 10, pp. 3486–3498, 2020.

[33] 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.

[34] 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.

[35] 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.

[36] V. A. Sindagi, R. Yasarla, and V. M. Patel, "Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method," *arXiv preprint arXiv:1910.12384*, 2019.

[37] Z. Yan, Y. Yuan, W. Zuo, X. Tan, Y. Wang, S. Wen, and E. Ding, "Perspective-guided convolution networks for crowd counting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 952–961.

[38] X. Liu, J. van de Weijer, and A. 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, pp. 7661–7669.

[39] D. B. Sam, N. N. Sajjan, H. Maurya, and R. V. Babu, "Almost unsupervised learning for dense crowd counting," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, vol. 27, 2019.

[40] G. Olmschenk, Z. Zhu, and H. Tang, "Generalizing semi-supervised generative adversarial networks to regression using feature contrasting," *Computer Vision and Image Understanding*, vol. 186, pp. 1–12, 2019.

[41] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International Conference on Machine Learning*, 2015, pp. 97–105.

[42] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 343–351.

[43] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 2208–2217.

[44] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.

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

[46] G. Ros, L. Sellart, J. Materzynska, D. Vázquez, and A. M. López, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3234–3243.

[47] 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.

[48] 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.

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

[50] J. Gao, Y. Yuan, and W. Qi, "Feature-aware adaptation and density alignment for crowd counting in video surveillance," *IEEE transactions on cybernetics*, pp. 1–12, 2020.

[51] A. Gonzalez-Garcia, J. van de Weijer, and Y. Bengio, "Image-to-image translation for cross-domain disentanglement," in *Advances in Neural Information Processing Systems*, 2018, pp. 1287–1298.

[52] W.-L. Chang, H.-P. Wang, W.-H. Peng, and W.-C. Chiu, "All about structure: Adapting structural information across domains for boosting semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1900–1909.

[53] B. Lu, J.-C. Chen, and R. Chellappa, "Unsupervised domain-specific deblurring via disentangled representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 225–10 234.

[54] Y.-J. Li, C.-S. Lin, Y.-B. Lin, and Y.-C. F. Wang, "Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation," *arXiv preprint arXiv:1909.09675*, 2019.

[55] L. Hu, M. Kan, S. Shan, and X. Chen, "Duplex generative adversarial network for unsupervised domain adaptation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1498–1507.

[56] Y. Cai, L. Chen, Z. Ma, C. Lu, C. Wang, and G. He, "Leveraging intra-domain knowledge to strengthen cross-domain crowd counting," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[57] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2794–2802.

[58] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[59] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[61] 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.

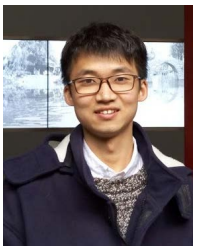
[62] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting." in *BMVC*, vol. 1, no. 2, 2012, p. 3.

[63] 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*, 2008, pp. 1–7.

[64] 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.

[65] 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.

[66] 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.
[67] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.



Junyu Gao received the B.E. degree and the Ph.D. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an, China, in 2015 and 2021, respectively. He is currently a researcher with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Tao Han received the B.E. degree in transportation equipment and control engineering from the Northwestern Polytechnical University, Xi'an, China, in 2019. he is currently working toward the M.S. degree in computer science and technology in the School of Artificial Intelligence, Optics and Electronics (iOPEN), 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 Artificial Intelligence, Optics and Electronics (iOPEN), 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 Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.