Error-Aware Density Isomorphism Reconstruction for Unsupervised Cross-Domain Crowd Counting

Yuhang He[1], Zhiheng Ma[1], Xing Wei[2], Xiaopeng Hong[3,4*], Wei Ke[2], Yihong Gong[2]
[1] College of Artificial Intelligence, Xi'an Jiaotong University
[2] School of Software Engineering, Xian Jiaotong University
[3]School of Cyber Science and Engineering, Xi'an Jiaotong University
[4]Research Center for Artificial Intelligence, Peng Cheng Laboratory
{hyh1379478,mazhiheng}@stu.xjtu.edu.cn, {weixing,hongxiaopeng, wei.ke,ygong}@mail.xjtu.edu.cn

Abstract

This paper focuses on the unsupervised domain adaptation problem for video-based crowd counting, in which we use labeled data as source domain and unlabelled video data as target domain. It is challenging as there is a huge gap between the source and the target domain and no annotations of samples are available in the target domain. The key issue is how to utilize unlabelled videos in the target domain for knowledge learning and transferring from the source domain. To tackle this problem, we propose a novel Error-aware Density Isomorphism REConstruction Network (EDIREC-Net) for cross-domain crowd counting. EDIREC-Net jointly transfers a pre-trained counting model to target domains using a density isomorphism reconstruction objective and models the reconstruction erroneousness by error reasoning. Specifically, as crowd flows in videos are consecutive, the density maps in adjacent frames turn out to be isomorphic. On this basis, we regard the density isomorphism reconstruction error as a self-supervised signal to transfer the pre-trained counting models to different target domains. Moreover, we leverage an estimation-reconstruction consistency to monitor the density reconstruction erroneousness and suppress unreliable density reconstructions during training. Experimental results on four benchmark datasets demonstrate the superiority of the proposed method and ablation studies investigate the efficiency and robustness. The source code is available at https://github.com/GehenHe/EDIREC-Net.

1 Introduction

Counting the number of targets in crowd scenarios, *i.e.*, the *crowd counting*, has drawn remarkable attention in recent years. It has a wide range of applications in video surveillance, traffic control, crowd behavior analysis, and so forth. Recently, benefited from the rapid development of Deep Convolutional Neural Network (DCNN) (He et al. 2016; Ke et al. 2020), DCNN based crowd counting methods (Zhang et al. 2019a; Liu et al. 2019; Guo et al. 2019; Bai et al. 2020; Ma et al. 2021) have achieved significant progresses and become the main stream of this field. Accurate and efficient as they are, these methods often rely on instance-level annotations and a massive number of training data to train the deep convolutional network. If the training data is scarce, these methods are easily prone to overfit the training data and may suffer performance deterioration when applied to other scenarios, (*i.e.*, target domain), in which data distributions are noticeably different from the training domain (*i.e.*, the source domain). This hampers their applications to real-world scenarios.

To mitigate the gap between the source domain and target domains, several semi- and un-supervised domain adaptation methods have been proposed for crowd counting. Semi-supervised methods (Change Loy, Gong, and Xiang 2013; Reddy et al. 2020) transfer the source-domain pre-trained counting model to target domains using a small number of labeled data from the target domain. These methods require data annotation in target domains, which are still expensive and laborious for practical usage. To address this problem, several unsupervised domain adaptation methods (Zhang et al. 2015; Wang, Li, and Xue 2019; Han et al. 2020) are proposed to transfer the pre-trained counting model to an unlabeled target domain by adversarial learning (Wang, Li, and Xue 2019; Han et al. 2020) or cross-domain image retrieval (Zhang et al. 2015). These unsupervised methods only focus on the image-level information, such as the image similarity (Zhang et al. 2015) and the domain distinctiveness (Wang, Li, and Xue 2019), but neglect the consistency information of videos sequences. As videos contain more information than static images, it is of great importance to study the problem of transferring pre-trained counting model to target domains using unlabeled videos.

In this paper, we focus on the cross-domain adaptation problem for video-based unsupervised crowd counting. The major challenge is how to explore concealed information in unlabeled videos for knowledge transfer from a source-domain to an unseen target domain. To address it, we propose a density isomorphism reconstruction objective for cross-domain knowledge transfer. As the crowd flow is smooth in a video sequence, the number of objects in a short period is consistent as well. It is thus reasonable to assume that the target distributions between adjacent frames are *isomorphic*, *i.e.*, mutually transformable using bijective map-

*Corresponding author

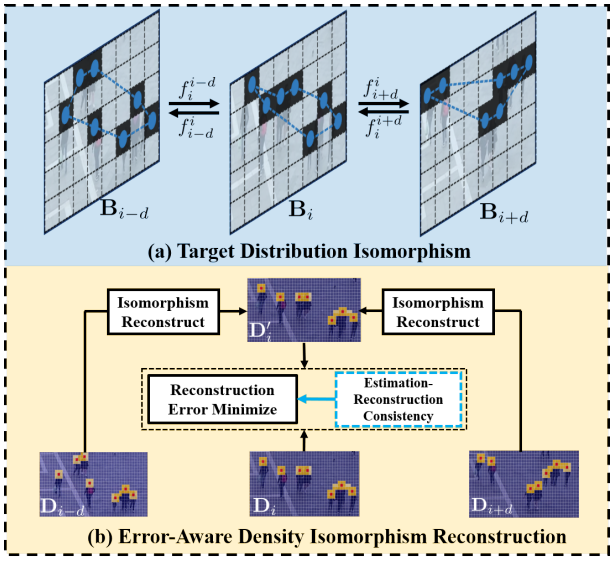

Figure 1: Illustration of (a) target distribution isomorphism and (b) error-aware density isomorphism reconstruction in the proposed EDIREC-Net. (a) By regarding each target as a point, we denote $\mathbf{B}_i \in \{0,1\}^{W\times H}$ as the target distribution map of image $\mathbf{I}_i$ at the i-th frame. The element of $\mathbf{B}_i$ indicates whether a target occurs at the corresponding location. Supposing the counting number in adjacent frames $\mathbf{I}_i$ and $\mathbf{I}_j$ are the same, $\mathbf{B}_i$ and $\mathbf{B}_j$ are isomorphic, *i.e.*, there is a mapping f_j^i from $\mathbf{B}_i$ to $\mathbf{B}_j$, which can be reversed by an inverse mapping f_i^j. (b) A density map $\mathbf{D}_i$ is *estimated* at time i, and $\mathbf{D}'_i$ is a *reconstructed* density map using $\mathbf{D}_{i-d}$ and $\mathbf{D}_{i+d}$. In the proposed EDIREC-Net, the difference between $\mathbf{D}'_i$ and $\mathbf{D}_i$ is minimized by considering the estimation-reconstruction consistency.

ping, as illustrated in Figure 1(a). This allows us to reconstruct density maps in the target domain by leveraging the connections between adjacent frames and regard the reconstruction error as a self-supervised signal to transfer the pre-trained counting model to the target domain.

However, it is still problematic to directly minimize the isomorphism reconstruction error, as both the estimated and reconstructed density maps may contain numerous errors. To reduce the impact of inaccurate density reconstruction in training, we propose to evaluate and monitor the density reconstruction erroneousness by *estimation-reconstruction consistency*. The intuition behind is that *only one truth prevails*. Specifically, if both the reconstruction and estimation are accurate and reliable, their outputs should be the same or close to equal; By contrary, if the differences between the outputs of reconstruction and estimation are non-negligible, their results are likely to be erroneous and thus unreliable. This inspires us to maximize the estimation-reconstruction consistency and monitor the erroneousness of density isomorphism reconstruction without annotations.

On this basis, we propose a novel end-to-end Error-aware Density Isomorphism REConstruction Network (EDIREC-Net) for unsupervised domain adaptation as illustrated in Figure 1 (b). EDIREC-Net simultaneously minimizes the density isomorphism reconstruction error and maximizes the estimation-reconstruction consistency. Experimental results compared with state-of-the-art methods on four benchmark datasets demonstrate the superiority of the EDIREC-Net and the ablation studies investigate the efficiency and robustness of the proposed method.

In summary, the main contributions of this paper include:

- We propose a novel end-to-end Error-aware Density Isomorphism Reconstruction Network for unsupervised cross-domain crowd counting in videos.
- We propose an error-aware density isomorphism reconstruction objective to transfer the pre-trained counting model to target domains.
- We develop a reconstruction erroneousness modeling mechanism to monitor the erroneousness of density reconstruction.

2 Related Work

2.1 Crowd Counting

In recent years, density map estimation based methods significantly improve the crowd counting performance (Lempitsky and Zisserman 2010; Zhang et al. 2019c; Wan and Chan 2019; Zhang et al. 2019b; Tan et al. 2019; **?**; Liu et al. 2020a; Liu, Yang, and Ding 2020; Ma et al. 2020). This kind of methods first estimate a density map for the input image and then calculate the count number by summing over the estimated density map. The method (Lempitsky and Zisserman 2010) first converts the target annotations (one labeled pixel for each target) into a ground-truth density map using Gaussian kernels, and then trains a density map estimator using a Maximum Excess over SubArrays (MESA) loss function. Ma *et al.* (Ma et al. 2019) propose a Bayesian Loss (BL) to construct a density distribution from the point annotations and adopt a count expectation supervision at each annotated point to train their density map estimator. The methods (Sindagi and Patel 2019; Chen, Su, and Wang 2020; Zhang et al. 2019a; Liu et al. 2019; Zhang et al. 2019b; Guo et al. 2019) integrate attention mechanism to crowd counting, which reduce the background noise by re-weighting target densities with attention mechanisms. Research studies (Cao et al. 2018; Shen et al. 2018; Idrees et al. 2013; Jiang et al. 2020; Chen et al. 2020) handle scale variation in crowd counting by generating multipolar normalized density maps (Xu et al. 2019), using a scale-aware pyramid network (Chen et al. 2020) or a scale aggregation network (Cao et al. 2018). Image-based crowd counting has been developed rapidly and video-based counting is also gaining attentions recently.

Crowd counting in videos. There are several approaches attempt to count the number of targets in video sequences (Xiong, Shi, and Yeung 2017; Fang et al. 2019; Liu et al. 2020b; Liu, Salzmann, and Fua 2019b). Xiong *et al.* (Xiong, Shi, and Yeung 2017) exploit a convolutional LSTM to capture both the spatial and temporal informations, and extend the ConvLSTM with a bidirectional LSTM to process the long-term information in crowd flows. The

method (Fang et al. 2019) proposes a locality-constrained spatial transformer network to exploit the spatial-temporal consistency in videos. Zhou *et al.* (Zou et al. 2019) introduce a deep trainable network to exploit the temporal dependencies in adjacent frames. Liu *et al.* (Liu, Salzmann, and Fua 2019b) exploit the consistency of people flows (PFlow) between adjacent images to improve the performance of crowd counting, and they use the optical flow to model the temporal correlation between consecutive image frames. Nevertheless, annotation in video is expensive and it's still unsolved how to transfer pre-trained counting model to unlabeled video sequences.

2.2 Domain Adaptation

When the training dataset is scarce, the performance of a pre-trained model may deteriorate dramatically when the target domain is different from the source domain. In the past decades, Domain Adaptation (DA) methods have been proposed to transfer the pre-trained model to target domains and improve the performance of the pre-trained model (Kang et al. 2019; Wang and Breckon 2019; Zhang and Davison 2019). According to whether requiring the annotating of target domain or not, these methods are further divided into two subcategories: Semi-supervised Domain Adaptation (SDA) and Unsupervised Domain Adaptation (UDA). SDA methods (Change Loy, Gong, and Xiang 2013; Reddy et al. 2020) transfer the pre-trained counting model to target domains by exploiting a small number of labeled data in the target domain. Change *et al.* (Change Loy, Gong, and Xiang 2013) develop a Semi-Supervised Regression (SSR) framework to exploit labeled data in the target domain to compensate the lacking of training data. Reddy *et al.* (Reddy et al. 2020) formulate the domain adaptation as a Few-Shot Scene Adaptation (FSSA) problem and propose a meta-learning based method to solve the problem.

Unsupervised domain adaptation for crowd counting. To get rid of the annotating of target domains, there are UDA crowd counting methods (Zhang et al. 2015; Wang, Li, and Xue 2019; Han et al. 2020) transferring a pre-trained model to target domains using unlabeled data. Zhang *et al.* (Zhang et al. 2015) propose a Crowd-Scene Crowd Counting (CSCC) method, which first retrievals images in the source domain that similar to the target domain and then fine-tunes the pre-trained counting model with these collected samples. Wang *et al.* (Wang, Li, and Xue 2019) propose a Count Objects via scale-aware adversarial Density Adaptation (CODA) algorithm for crowd counting, which introduces an adversarial training approach to adapt the target domain. During training, they take both the labeled source domain data and unlabeled target domain data as input, and introduce a discriminator for distinguishing whether the density map is generated from the source or the target domain. Han *et al.* (Han et al. 2020) propose a Semantic Consistency Predictor (SCP) to estimate crowd masks, which eliminates the background clutter and train the counting model using an adversarial learning framework. However, the consistency information of videos sequences is neglected.

3 Methodology

In this section, we first overview the framework of the proposed method and then provide detailed descriptions of our key techniques.

3.1 General Framework

Given an unlabeled image sequence $\mathcal{I} = \{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_T\}$ of length T collected from the target domain. Let $\varphi(\cdot;\Theta)$ be a density map estimator with parameter set Θ. For each input image $\mathbf{I} \in \mathbb{R}^{W_I \times H_I}$, the $\varphi(\cdot;\Theta)$ estimates a density map $\mathbf{D} \in \mathbb{R}_{\geq 0}^{W_D \times H_D}$ for the input image:

$$\mathbf{D} = \varphi(\mathbf{I};\Theta), \tag{1}$$

where W_I and H_I (W_D and H_D) are the width and height of the image frame (density map), respectively, and each element $\mathbf{D}(u,v)$ in $\mathbf{D}$ denotes the number of targets at position (u,v). We denote $\varphi_s(\cdot;\Theta_s)$ as the pre-trained counting model and $\varphi_t(\cdot;\Theta_t)$ the target-domain counting model.

The objective of EDIREC-Net is to transfer the prior knowledge of $\varphi_s(\cdot;\Theta_s)$ to the target domain, and train a target-domain model $\varphi_t(\cdot;\Theta_t)$ using $\mathcal{I}$ that achieves more accurate counting results on the target domain. As illustrated in Figure 2, the proposed EDIREC-Net contains three major modules: 1) density and erroneousness inference module, 2) isomorphism reconstruction module and 3) reconstruction erroneousness modeling module. During training, Module 1 takes an image tuple $\mathcal{I}_i^d = \{\mathbf{I}_{i-d}, \mathbf{I}_i, \mathbf{I}_{i+d}\}$ as input (d is a time interval) and infers a density map $\mathbf{D}_j \in \mathbb{R}_{\geq 0}^{W_D \times H_D}$ and a reconstruction erroneousness matrix $\mathbf{E}_j \in \mathbb{R}_{\geq 0}^{W_D \times H_D}$ for each image $\mathbf{I}_j \in \mathcal{I}_i^d$ using the density map header Φ_D and reconstruction erroneousness header Φ_E, respectively. Module 2 takes the estimated density maps as input and reconstructs the density map of image $\mathbf{I}_i$ using $\mathbf{D}_{i-d}$ and $\mathbf{D}_{i+d}$ by isomorphism reconstruction. It then uses an error-aware density isomorphism reconstruction objective $\mathcal{L}_{iso}$ to optimize the network parameter, which simultaneously minimizes the density reconstruction error and suppresses unreliable density reconstruction during training. Module 3 further constrains the erroneousness of isomorphism reconstruction using a regularization term $\mathcal{L}_{mod}$.

Based on the above descriptions, the objective function of the EDIREC-Net can be written as:

$$\min \sum_{i=1}^{T} \Big[\mathcal{L}_{iso}(\mathcal{I}_i^d) + \mathcal{L}_{mod}(\mathcal{I}_i^d) \Big], \tag{2}$$

where $\mathcal{L}_{iso}(\cdot)$ is the proposed error-aware density isomorphism reconstruction objective and $\mathcal{L}_{mod}(\cdot)$ is a regularization term to models the reconstruction erroneousness.

It is worth mentioning that, to avoid the training of $\varphi_t(\cdot;\Theta_t)$ being stuck into trivial solutions (such as all-zero outputs), we employ an "anchor" counting model $\varphi_t(\cdot;\Theta_a)$ to infer the density map and erroneousness matrix of the $\mathbf{I}_i$ in $\mathcal{I}_i^d$. The network architecture and parameter initialization of $\varphi_t(\cdot;\Theta_a)$ are identical to the ones of $\varphi_t(:,\Theta_t)$ while the parameter set Θ_a is updated using an exponential moving average (Tarvainen and Valpola 2017). More implementation details are provided in Section 4.2.

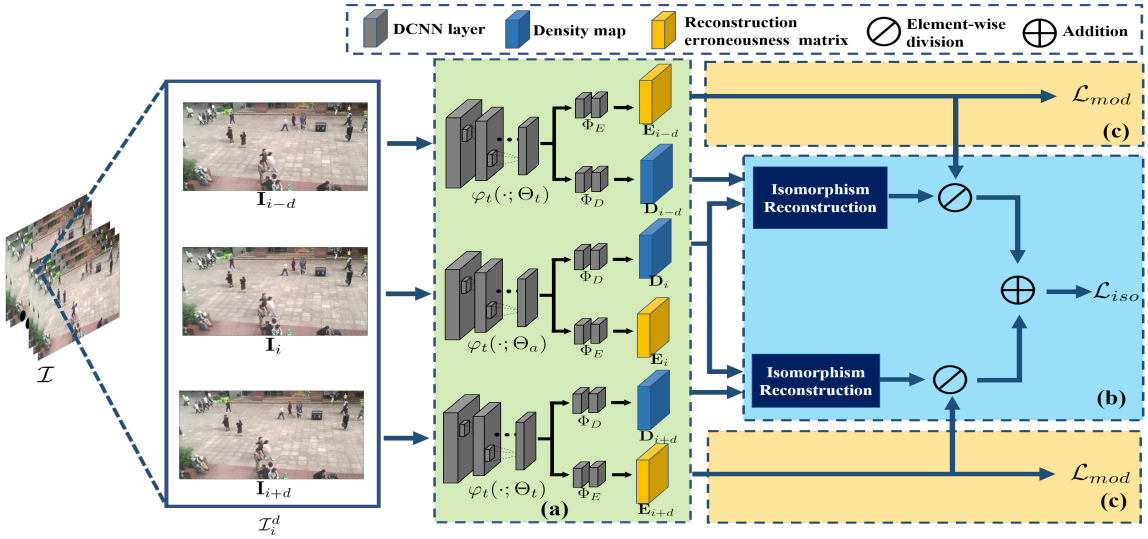

Figure 2: Framework overview of the EDIREC-Net. The input to the framework is an unlabeled image sequence $\mathcal{I}$ collected from the target domain. The proposed method contains three major modules: (a) the density and erroneousness inference module, (b) the isomorphism reconstruction module and (c) the reconstruction erroneousness modeling module.

3.2 Density Isomorphism Reconstruction

Given the estimated density maps $\mathbf{D}_{i-d}$, $\mathbf{D}_i$ and $\mathbf{D}_{i+d}$ of an image tuple $\mathcal{I}_i^d = \{\mathbf{I}_{i-d}, \mathbf{I}_i, \mathbf{I}_{i+d}\}$, respectively. We aim to reconstruct the density map of image $\mathbf{I}_i$ using $\mathbf{D}_{i-d}$ and $\mathbf{D}_{i+d}$ and regard the reconstruction error as a self-supervised signal to transfer the pre-trained model.

The key of the density isomorphism reconstruction is how to find a mapping correspondence between each pair of density maps. A straightforward solution is to minimizing an earth mover's distance (EMD) (Levina and Bickel 2001) and transporting a density map to an another one by calculating a density transportation correspondence. However, as the estimated density maps are inaccurate and contain errors, these mapping correspondences are defectives and could be deleterious to the training of target-domain model.

To tackle this problem, instead of calculating mapping correspondence using density maps directly, we first calculate an image mapping correspondence using image frames and then convert the image-level correspondence to a density one by matrix transformation. Let $\mathbf{M}_i^j \in \mathbb{R}^{W_I \times H_I \times 2}$ be an image mapping matrix from $\mathbf{I}_j$ to $\mathbf{I}_i$ and $\mathbf{G}_i^j \in \mathbb{R}^{W_D \times H_D \times 2}$ be a density mapping matrix from $\mathbf{D}_j$ to $\mathbf{D}_i$. Given images $\mathbf{I}_i$ and $\mathbf{I}_j$, we compute a mapping matrix $\mathbf{M}_i^j$ from $\mathbf{I}_j$ to $\mathbf{I}_i$ by minimizing an image reconstruction error:

$$\mathbf{M}_i^{j*} = \underset{\mathbf{M}_i^j}{\operatorname{argmin}} \|\mathbf{I}_i - \rho(\mathbf{I}_j, \mathbf{M}_i^j)\|^2, \tag{3}$$

where $\rho(\cdot,\cdot)$ is a warping function. In this paper, as $\mathbf{I}_{i-d}$, $\mathbf{I}_i$ and $\mathbf{I}_{i+d}$ are consecutive, each pixel $\mathbf{I}_i(u,v)$ in $\mathbf{I}_i$ can be approximated by quadratic polynomials of neighbor pixels of $\mathbf{I}_{i-d}(u,v)$ and $\mathbf{I}_{i+d}(u,v)$, respectively. Thus, the image mapping matrix $\mathbf{M}_i^j$ can be efficiently computed using the Gunner-Farneback algorithm (Farneback 2003). We denote by $\mathbf{M}_i^{i-d}$ and $\mathbf{M}_i^{i+d}$ the computed image mapping matrix from $\mathbf{I}_{i-d}$ and $\mathbf{I}_{i+d}$ to $\mathbf{I}_i$, respectively.

As a density map $\mathbf{D}_i$ (or $\mathbf{D}_j$) can yet be regarded as a downsampled target distribution of $\mathbf{I}_i$ (or $\mathbf{I}_j$) (Lempitsky and Zisserman 2010), the mapping correspondence from $\mathbf{D}_j$ to $\mathbf{D}_i$ and the one from $\mathbf{I}_j$ to $\mathbf{I}_i$ are linearly correlated. So that the density mapping matrix $\mathbf{G}_i^j$ can be obtained from $\mathbf{M}_i^j$ by linear sampling and scaling, and the density mapping matrices $\mathbf{G}_i^{i-d}$ and $\mathbf{G}_i^{i+d}$ from $\mathbf{D}_{i-d}$ and $\mathbf{D}_{i+d}$ to $\mathbf{D}_i$ can be obtained by:

$$\mathbf{G}_i^{i-d}(u,v) = \mathbf{M}_i^{i-d}\left(\tfrac{W_I}{W_D}u, \tfrac{H_I}{H_D}v\right) \cdot \sqrt{\tfrac{W_D^2+H_D^2}{W_I^2+H_I^2}}, \tag{4}$$

$$\mathbf{G}_i^{i+d}(u,v) = \mathbf{M}_i^{i+d}\left(\tfrac{W_I}{W_D}u, \tfrac{H_I}{H_D}v\right) \cdot \sqrt{\tfrac{W_D^2+H_D^2}{W_I^2+H_I^2}}. \tag{5}$$

On this basis, we can reconstruct the density map of image $\mathbf{I}_i$ using $\mathbf{D}_{i-d}$ and $\mathbf{D}_{i+d}$:

$$\mathbf{D}_i^{i-d'}(x,y) = \mathbf{D}_{i-d}(u,v), \forall (x,y) = \mathbf{G}_i^{i-d}(u,v), \tag{6}$$

$$\mathbf{D}_i^{i+d'}(x,y) = \mathbf{D}_{i+d}(u,v), \forall (x,y) = \mathbf{G}_i^{i+d}(u,v), \tag{7}$$

where $\mathbf{D}_i^{i-d'}$ and $\mathbf{D}_i^{i+d'}$ denotes the reconstructed density map of image $\mathbf{I}_i$ using $\mathbf{D}_{i-d}$ and $\mathbf{D}_{i+d}$, respectively.

3.3 Reconstruction Erroneousness Modeling

To monitor the erroneousness of density reconstruction and suppress inaccurate density reconstructions during training, we leverage the estimation-reconstruction consistency to model the reconstruction erroneousness.

Let $\mathbf{E}_i \in \mathbb{R}_{\geq 0}^{W_D \times H_D}$ be a reconstruction erroneousness matrix of $\mathbf{D}_i$. The larger the reconstruction erroneousness of $\mathbf{D}_i(u,v)$ is, the higher $\mathbf{E}_i(u,v)$ should be assigned, and vice versa. As there are no "ground truth" density maps provided in $\mathcal{I}_i^d$, evaluating the correctness of the density reconstruction is difficult. To settle this problem, we propose to

evaluate the erroneousness of density reconstruction according to the estimation-reconstruction consistency. Specifically, let $\mathbf{D}_i'$ be a reconstructed density map of $\mathbf{I}_i$. When the reconstructed density map $\mathbf{D}_i'$ and the estimated density map $\mathbf{D}_i$ are consistent at (u, v), *i.e.*, $|\mathbf{D}_i(u,v) - \mathbf{D}_i'(u,v)| \to 0$, it is very likely that both $\mathbf{D}_i(u,v)$ and $\mathbf{D}_i'(u,v)$ are correct. By contrary, if $\mathbf{D}_i'$ and $\mathbf{D}_i$ are inconsistent at (u, v), *i.e.*, $|\mathbf{D}_i(u,v) - \mathbf{D}_i'(u,v)| \gg 0$, $\mathbf{D}_i(u,v)$ and/or $\mathbf{D}_i'(u,v)$ are very likely to be erroneous and thus unreliable. Therefore, the larger $|\mathbf{D}_i(u,v) - \mathbf{D}_i'(u,v)|$ is, the larger reconstruction erroneousness $\mathbf{E}_i(u,v)$ should be assigned.

Embedding this objective to the training of the EDIREC-Net, we derive an error-aware density isomorphism reconstruction objective, which can be formulated as:

$$\mathcal{L}_{iso}(\mathcal{I}_i^d) = \left\| \left| \mathbf{D}_i - \mathbf{D}_i^{i-d'} \right|_e \oslash \mathbf{E}_{i-d} \right\|^2 + \left\| \left| \mathbf{D}_i - \mathbf{D}_i^{i+d'} \right|_e \oslash \mathbf{E}_{i+d} \right\|^2, \tag{8}$$

where $\mathbf{D}_i^{i-d'}$ and $\mathbf{D}_i^{i+d'}$ are the reconstructed density maps using Eqs.(6) and (7), $|\cdot|_e$ outputs the element-wise $L2$-norm of the input matrix and $\oslash$ denotes element-wise matrix division. Nevertheless, solely minimizing $\mathcal{L}_{iso}$ may lead to trivial solutions and the erroneousness matrices $\mathbf{E}_{i\pm d}$ could be extraordinary large. To solve this problem, we develop a regularization term $\mathcal{L}_{mod}(\cdot)$ to constrain the values of $\mathbf{E}_{i\pm d}$:

$$\mathcal{L}_{mod}(\mathcal{I}_i^d) = \log(\mathbf{E}_{i-d}) + \log(\mathbf{E}_{i+d}), \tag{9}$$

where $\log(\cdot)$ outputs the logarithmic sum of the input matrix.

On this basis, for each image tuple $\mathcal{I}_i^d$, the objective function of the proposed method can be written as:

$$\mathcal{L}(\mathcal{I}_i^d) = \mathcal{L}_{iso}(\mathcal{I}_i^d) + \mathcal{L}_{mod}(\mathcal{I}_i^d), \tag{10}$$

where $\mathcal{L}_{iso}$ encourages the network minimizing the difference between $\mathbf{D}_i$ and $\mathbf{D}_i^{i\pm d'}$, and $\mathcal{L}_{iso}$ and $\mathcal{L}_{mod}$ jointly models the erroneousness matrices $\mathbf{E}_{i\pm d}$ to suppress the impact of unreliable density reconstructions during training.

4 Experiment

4.1 Dataset

To investigate the effectiveness of the proposed method in cross-domain crowd counting, we conduct domain adaptation experiments from the source domain datasets (*i.e.*, the UCF-QNRF (Idrees et al. 2018) dataset) to different target domain datasets (*i.e.*, UCSD, MALL, VENICE, FDST), where the target domain datasets are video sequences collected from real-world scenarios. A brief description of these datasets is provided as the following:

UCF-QNRF (Idrees et al. 2018). The dataset is a large-scale dataset containing 1535 images with 1,251,642 point annotations, where each image includes about 800 targets in average. The dataset has 1,201 images for training and the remaining 334 images are used for testing.

UCSD (Chan, Liang, and Vasconcelos 2008). It contains 2,000 frames captured from a surveillance camera. The resolution of images frames is 238×158 and the frame rate is 10 fps. Following the experiment setting in (Chan, Liang, and Vasconcelos 2008), we use the frames from 601 to 1400 for training and the remaining 1200 frames for testing.

MALL (Chen et al. 2012). This dataset contains 2,000 frames captured from a mall with a fixed resolution 640×480. The video frame rate is about 2 fps and there are about 30 targets in each frame. Following the settings in (Chen et al. 2012), we use the first 800 frames for training and keep the remaining 1,200 frames for testing.

VENICE (Liu, Salzmann, and Fua 2019a). It contains 167 annotated frames from 4 different scenarios. The image resolution is 1280×720 and there are about 250 targets in each frame. Following (Liu, Salzmann, and Fua 2019a), we use 80 images from a single scenario for training and keep the rest images from the other 3 scenarios for testing.

FDST (Fang et al. 2019). This dataset captures 100 video sequences from 13 different scenes with 150,000 image frames and 394,081 annotated head points. There are 60 video sequences re used for training and the rest are used for testing. The videos are captured at 30 fps with a resolution of 1920×1080.

4.2 Implementation Details

Network Architecture. We use the VGG-19 architecture (Simonyan and Zisserman 2014) as the backbone of our density map estimator, where the fully-connected layers are removed. We then feed the output feature map to a density estimation header (Φ_D) and a reconstruction erroneousness prediction header (Φ_E), respectively, which both consist of two 3×3 and one 1×1 convolutional layers.

Training Details. The source-domain density map estimator $\varphi_s(\cdot; \Theta_s)$ is pre-trained on the UCF-QNRF dataset using the Bayesian Loss (BL) (Ma et al. 2019).

The backbone and the density map header of $\varphi_t(\cdot; \Theta_t)$ are initialized using Θ_s and the erroneousness estimation header is randomly initialized. The "anchor" counting model $\varphi_t(\cdot; \Theta_a)$ is identical to $\varphi_t(\cdot; \Theta_t)$ (including the network architecture and parameter initialization) except the parameter updating method. During training, Θ_t is updated using an Adam optimizer (Kingma and Ba 2014) with a learning rate of 10^{-5}, while Θ_a is updated using an exponential moving average (Tarvainen and Valpola 2017): $\Theta_a = \alpha\Theta_a + \Theta_t$, where α is the parameter of moving step. In this paper, we fix $\alpha = 0.999$, which means Θ_a is about the weight average of the latest $1/(1-\alpha) = 1000$ iterations of Θ_t. The time interval parameter d is fix to $d = 3$ according to the experimental results in Section 4.5.

4.3 Evaluation Metrics

We adopt two widely used crowd counting metrics to evaluate the proposed method: Mean Absolute Error (MAE) and Mean Squared Error (MSE), which are defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|C_i^{gt} - C_i|, \tag{11}$$

$$MSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}|C_i^{gt} - C_i|^2}, \tag{12}$$

Supervision	Method	Venice		UCSD		MALL		FDST	
		MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓
—	Baseline	33.95	39.44	7.96	8.54	4.27	5.94	4.77	8.33
Supervised	PFlow	15.00	19.60	*0.81*	*1.07*	—	—	2.84	3.57
	BL	*9.99*	*14.24*	0.84	1.08	*1.54*	*2.00*	*1.42*	*1.88*
Semi-supervised	SSR	19.84	31.13	1.68	2.07	2.69	3.38	5.41	6.13
	FSSA	17.83	25.24	1.45	1.85	2.32	2.97	2.96	3.86
Unsupervised	CSCC	18.05	22.34	8.89	9.87	4.01	4.99	5.15	7.84
	CODA	31.39	37.17	5.25	6.07	3.37	4.43	4.74	8.27
	SCP	22.79	26.52	4.55	5.71	3.03	4.04	4.28	6.74
	Ours-*w/o mod*	14.66	17.48	2.22	2.71	3.17	4.03	3.97	4.76
	Ours	**11.23**	**15.16**	**1.79**	**2.47**	**2.36**	**3.12**	**3.25**	**3.94**

Table 1: Performance Evaluation on Four Benchmark Datasets.

where N is the number of testing images, C_t^{gt} and C_t are the ground-truth and the estimated count number of the t-th frame, respectively.

4.4 Experimental Evaluations

In Table 1, we compare the proposed method with both semi- and un-supervised domain adaptation methods for crowd counting (described in Section 2). Moreover, we also compare the proposed method with representative supervised methods, *i.e.*, PFlow (Liu, Salzmann, and Fua 2019b) and BL (Ma et al. 2019). The best results of the supervised, semi-supervised and unsupervised methods are denoted by *italics*, underline and **bold**, respectively. The *Baseline* method outputs the results of the pre-trained model. The *Ours* method outputs the counting results of the proposed method and the *Ours-w/o mod* outputs the results without the reconstruction erroneousness modeling. For fair comparison, all the methods we compared use the pre-trained model as the proposed method. From Table 1, we make the following important observations:

- The performance of the pre-trained (*i.e.*, the *Baseline*) is inferior on target domains due to the domain gap between targets domains and the source domain.
- Without requiring any additional annotations, the proposed method significantly improves the performance of the *Baseline* method on the target domain datasets (by reducing 67%, 77%, 42%, 28% MAE error on the Venice, UCSD, MALL, FDST datasets, respectively).
- Compared with *Ours-w/o mod*, the *Ours* method steadily improves the counting performance (around 20% improvement on the four benchmark datasets) by modeling the reconstruction erroneousness.
- The proposed method outperforms all the unsupervised cross domain counting methods and achieves highly competitive results with state-of-the-art semi-supervised methods.
- The proposed method achieves comparable results with the representative fully-supervised methods (upper bounds of the domain adaptation methods).

The key to achieve such results is that, the EDIREC-Net regards the density isomorphism reconstruction error as a self-supervised signal, which is efficient to transfer the pre-trained counting models to different target domains. Besides, the erroneousness modeling mechanism provides an opportunity to monitor the reliability of isomorphism reconstruction, which can further improve the counting performance. To get more insights, we illustrate counting examples of different methods on the four target datasets in Figure 4. We can see that: 1) the pre-trained counting model (*Baseline*) produces inaccurate counting results such as missing counting (the yellow circles) and redundant counting (the red circles) in target domains. 2) the *Ours-w/o mod* transfers the pre-trained model to target domains and significantly improves the counting performance. 3) By exploiting the reconstruction erroneousness modeling mechanism, the *Ours* achieves more accurate counting results.

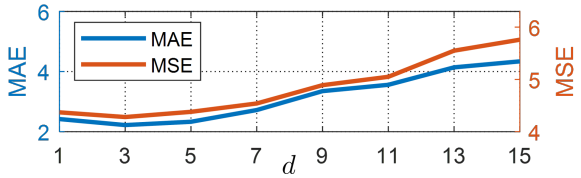

Figure 3: Influence of different d values.

4.5 Ablation Studies

Influence of d To study the influence of d, we conduct experiments on the validations sets (100 images randomly sampled from the training set) of the MALL dataset. The meta-parameter d affects the collection of the image tuple $\mathcal{I}_i^d = \{\mathbf{I}_{i-d}, \mathbf{I}_i, \mathbf{I}_{i+d}\}$. It can be seen from Figure 3 that, when d is smaller than 5, the accuracy keeps steady and the best performance is achieved at $d = 3$. When d exceeds 5, the counting performance start to decrease. This is mainly because when d is too large, the number of targets in $\mathcal{I}_i^d$ are different and the isomorphism of density maps are collapsed.

Robustness to Different Pre-trained Models. To study the robustness of the proposed method to different baseline methods and source domain datasets, we adopt the MESA (Lempitsky and Zisserman 2010) as an additional baseline method and the ShanghaiTech-A (Zhang et al. 2016) dataset as an additional source domain dataset. The ShanghaiTech-A includes 300 training images and 182 testing images with about 500 targets per image. We denote by

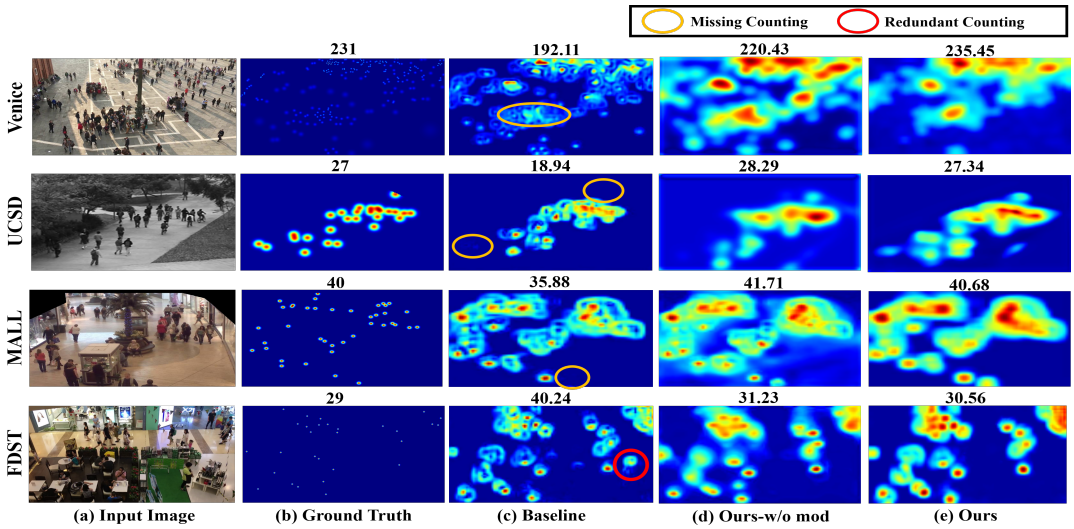

Figure 4: Counting examples of the (b) Ground-Truth, (c) Baseline, (d) Our-w/o mod and (e) Ours methods. The warmer color denotes the higher density. The number upon each density map denotes the predicted count number, and the yellow and red circles mark the missing and redundant counting of the Baseline methods, respectively.

Source	Method	Venice		UCSD		MALL		FDST	
		MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓
ShanghaiTech-A	Baseline-MESA	51.57	53.68	16.80	17.81	15.67	16.75	12.80	25.59
	Ours	17.83	22.19	5.13	5.83	5.57	6.54	6.27	7.64
	Baseline-BL	40.13	51.54	15.36	16.18	12.48	12.99	5.01	8.09
	Ours	14.10	19.13	4.22	5.01	4.77	5.93	3.96	5.12
UCF-QNRF	Baseline-MESA	43.16	57.88	9.04	9.77	5.71	6.67	6.12	7.57
	Ours	13.05	15.72	2.64	3.60	4.65	6.01	4.95	6.10
	Baseline-BL	33.95	39.44	7.96	8.54	4.27	5.94	4.77	8.33
	Ours	**11.23**	**15.16**	**1.79**	**2.47**	**2.36**	**3.12**	**3.25**	**3.94**

Table 2: Robustness to Different Pre-trained Models.

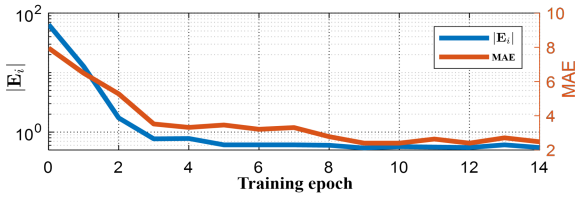

Figure 5: $|\mathbf{E}_i|$ and MAE curves during training.

Baseline-BL and *Baseline-MESA* the pre-trained model using the Bayesian Loss (Ma et al. 2019) and the MESA Loss, respectively. From Table 2, we can see that: 1) as the size of UCF-QNRF is much larger than the one of ShanghaiTech-A (1535 v.s. 300), the UCF-QNRF pre-trained model is more accurate and robust than the ShanghaiTech-A one. 2) Even the pre-trained model is not very accurate, the proposed method can steadily and significantly improve the performance (more than 30% MAE improvement) and achieves favorable counting results on all the four datasets.

Reconstruction Erroneousness Analysis To analysis the reconstruction erroneousness modeling, we draw the curves of $|\mathbf{E}_i|$ and MAE error on the validation set of UCSD dataset in Figure 5. We can see that, at the beginning of training, the estimated density maps are not very accurate and the density reconstruction erroneousness $|\mathbf{E}_i|$ is high. By transferring the pre-trained counting model to the target domain, the reconstruction erroneousness $|\mathbf{E}_i|$ decreases significantly. Meanwhile, the counting performance improves steadily when the reconstruction erroneousness decreasing. This demonstrates the erroneousness modeling mechanism can efficiently suppress the reconstruction erroneousness and improve the counting performance.

5 Conclusion

In this paper, we propose an Error-aware Density Isomorphism REConstruction Network (EDIREC-Net) to transfer the source-domain prior knowledge to target domains using unlabeled video sequences. We transfer the pre-trained counting model to target domains using the density isomorphism reconstruction objective and develop a reconstruction erroneousness modeling mechanism to monitor the erroneousness of density reconstructions. Experimental results on four benchmark datasets demonstrate the superiority of the proposed method.

Acknowledgements

This work is funded by National Key Research and Development Project of China under Grant No.2019YFB1312000 and 2020AAA0105600, National Natural Science Foundation of China under Grant No. 62006183, 62076195, and 62006182, and by China Postdoctoral Science Foundation under Grant No. 2020M683489.

References

Bai, S.; He, Z.; Qiao, Y.; Hu, H.; Wu, W.; and Yan, J. 2020. Adaptive Dilated Network With Self-Correction Supervision for Counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4594–4603.

Cao, X.; Wang, Z.; Zhao, Y.; and Su, F. 2018. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 734–750.

Chan, A. B.; Liang, Z.-S. J.; and Vasconcelos, N. 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–7. IEEE.

Change Loy, C.; Gong, S.; and Xiang, T. 2013. From semi-supervised to transfer counting of crowds. In *Proceedings of the IEEE International Conference on Computer Vision*, 2256–2263.

Chen, J.; Su, W.; and Wang, Z. 2020. Crowd counting with crowd attention convolutional neural network. *Neurocomputing* 382: 210–220.

Chen, K.; Loy, C. C.; Gong, S.; and Xiang, T. 2012. Feature mining for localised crowd counting. In *BMVC*, volume 1, 3.

Chen, Y.; Gao, C.; Su, Z.; He, X.; and Liu, N. 2020. Scale-Aware Rolling Fusion Network for Crowd Counting. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Fang, Y.; Zhan, B.; Cai, W.; Gao, S.; and Hu, B. 2019. Locality-constrained spatial transformer network for video crowd counting. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 814–819. IEEE.

Farneback, G. 2003. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, 363–370. Springer.

Guo, D.; Li, K.; Zha, Z.-J.; and Wang, M. 2019. Dadnet: Dilated-attention-deformable convnet for crowd counting. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1823–1832.

Han, T.; Gao, J.; Yuan, Y.; and Wang, Q. 2020. Focus on Semantic Consistency for Cross-domain Crowd Understanding. *arXiv preprint arXiv:2002.08623* .

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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

Idrees, H.; Tayyab, M.; Athrey, K.; Zhang, D.; Al-Maadeed, S.; Rajpoot, N.; and Shah, M. 2018. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 532–546.

Jiang, X.; Zhang, L.; Xu, M.; Zhang, T.; Lv, P.; Zhou, B.; Yang, X.; and Pang, Y. 2020. Attention Scaling for Crowd Counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4706–4715.

Kang, G.; Jiang, L.; Yang, Y.; and Hauptmann, A. G. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4893–4902.

Ke, W.; Zhang, T.; Huang, Z.; Ye, Q.; Liu, J.; and Huang, D. 2020. Multiple Anchor Learning for Visual Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10206–10215.

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

Lempitsky, V.; and Zisserman, A. 2010. Learning to count objects in images. In *Advances in neural information processing systems*, 1324–1332.

Levina, E.; and Bickel, P. 2001. The earth mover's distance is the mallows distance: Some insights from statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, 251–256. IEEE.

Liu, L.; Chen, J.; Wu, H.; Chen, T.; Li, G.; and Lin, L. 2020a. Efficient Crowd Counting via Structured Knowledge Transfer. *arXiv preprint arXiv:2003.10120* .

Liu, L.; Lu, H.; Zou, H.; Xiong, H.; Cao, Z.; and Shen, C. 2020b. Weighing Counts: Sequential Crowd Counting by Reinforcement Learning. *arXiv preprint arXiv:2007.08260* .

Liu, N.; Long, Y.; Zou, C.; Niu, Q.; Pan, L.; and Wu, H. 2019. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3225–3234.

Liu, W.; Salzmann, M.; and Fua, P. 2019a. Context-aware crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5099–5108.

Liu, W.; Salzmann, M.; and Fua, P. 2019b. Estimating People Flows to Better Count them in Crowded Scenes. *arXiv preprint arXiv:1911.10782* .

Liu, X.; Yang, J.; and Ding, W. 2020. Adaptive Mixture Regression Network with Local Counting Map for Crowd Counting. *arXiv preprint arXiv:2005.05776* .

Ma, Z.; Wei, X.; Hong, X.; and Gong, Y. 2019. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, 6142–6151.

Ma, Z.; Wei, X.; Hong, X.; and Gong, Y. 2020. Learning Scales from Points: A Scale-Aware Probabilistic Model for Crowd Counting. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 220228. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885. doi:10.1145/3394171.3413642. URL https://doi.org/10.1145/3394171.3413642.

Ma, Z.; Wei, X.; Hong, X.; Lin, H.; Qiu, Y.; and Gong, Y. 2021. Learning to Count via Unbalanced Optimal Transport. *Proceedings of the AAAI Conference on Artificial Intelligence* .

Reddy, M. K. K.; Hossain, M.; Rochan, M.; and Wang, Y. 2020. Few-Shot Scene Adaptive Crowd Counting Using Meta-Learning. In *The IEEE Winter Conference on Applications of Computer Vision*, 2814–2823.

Shen, Z.; Xu, Y.; Ni, B.; Wang, M.; Hu, J.; and Yang, X. 2018. Crowd counting via adversarial cross-scale consistency pursuit. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5245–5254.

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

Sindagi, V. A.; and Patel, V. M. 2019. Ha-ccn: Hierarchical attention-based crowd counting network. *IEEE Transactions on Image Processing* 29: 323–335.

Tan, X.; Tao, C.; Ren, T.; Tang, J.; and Wu, G. 2019. Crowd Counting via Multi-layer Regression. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1907–1915.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, 1195–1204.

Wan, J.; and Chan, A. 2019. Adaptive density map generation for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, 1130–1139.

Wang, L.; Li, Y.; and Xue, X. 2019. CODA: Counting Objects via Scale-aware Adversarial Density Adaption. *arXiv preprint arXiv:1903.10442* .

Wang, Q.; and Breckon, T. P. 2019. Unsupervised Domain Adaptation via Structured Prediction Based Selective Pseudo-Labeling. *arXiv preprint arXiv:1911.07982* .

Xiong, F.; Shi, X.; and Yeung, D.-Y. 2017. Spatiotemporal modeling for crowd counting in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 5151–5159.

Xu, C.; Qiu, K.; Fu, J.; Bai, S.; Xu, Y.; and Bai, X. 2019. Learn to Scale: Generating Multipolar Normalized Density Maps for Crowd Counting. In *Proceedings of the IEEE International Conference on Computer Vision*, 8382–8390.

Zhang, A.; Shen, J.; Xiao, Z.; Zhu, F.; Zhen, X.; Cao, X.; and Shao, L. 2019a. Relational attention network for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, 6788–6797.

Zhang, A.; Yue, L.; Shen, J.; Zhu, F.; Zhen, X.; Cao, X.; and Shao, L. 2019b. Attentional neural fields for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, 5714–5723.

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

Zhang, L.; Shi, Z.; Cheng, M.-M.; Liu, Y.; Bian, J.-W.; Zhou, J. T.; Zheng, G.; and Zeng, Z. 2019c. Nonlinear regression via deep negative correlation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Zhang, Y.; and Davison, B. D. 2019. Modified Distribution Alignment for Domain Adaptation with PretrainedInception ResNet. *arXiv preprint arXiv:1904.02322* .

Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; and Ma, Y. 2016. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 589–597.

Zou, Z.; Shao, H.; Qu, X.; Wei, W.; and Zhou, P. 2019. Enhanced 3D convolutional networks for crowd counting. *arXiv preprint arXiv:1908.04121* .