Explicit Invariant Feature Induced Cross-Domain Crowd Counting

Yiqing Cai[1*], Lianggangxu Chen[1*], Haoyue Guan[3], Shaohui Lin[1†], Changhong Lu[2], Changbo Wang[1†], Gaoqi He[1, 4, 5†]

[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
[2]School of Mathematical Sciences, East China Normal University, Shanghai, China
[3]Johns Hopkins University, Mason Hall, USA
[4]Innovation Center for AI and Drug Discovery, East China Normal University, Shanghai, China
[5]Chongqing Key Laboratory of Precision Optics, Chongqing Institute of East China Normal University, Chongqing, China
{yqcai, lgxchen}@stu.ecnu.edu.cn, ghyherbert@163.com, shlin@cs.ecnu.edu.cn
chlu@math.ecnu.edu.cn, {cbwang, gqhe}@cs.ecnu.edu.cn

Abstract

Cross-domain crowd counting has shown progressively improved performance. However, most methods fail to explicitly consider the transferability of different features between source and target domains. In this paper, we propose an innovative explicit Invariant Feature induced Cross-domain Knowledge Transformation framework to address the inconsistent domain-invariant features of different domains. The main idea is to explicitly extract domain-invariant features from both source and target domains, which builds a bridge to transfer more rich knowledge between two domains. The framework consists of three parts, global feature decoupling (GFD), relation exploration and alignment (REA), and graph-guided knowledge enhancement (GKE). In the GFD module, domain-invariant features are efficiently decoupled from domain-specific ones in two domains, which allows the model to distinguish crowds features from backgrounds in the complex scenes. In the REA module both inter-domain relation graph (Inter-RG) and intra-domain relation graph (Intra-RG) are built. Specifically, Inter-RG aggregates multi-scale domain-invariant features between two domains and further aligns local-level invariant features. Intra-RG preserves task-related specific information to assist the domain alignment. Furthermore, GKE strategy models the confidence of pseudo-labels to further enhance the adaptability of the target domain. Various experiments show our method achieves state-of-the-art performance on the standard benchmarks. Code is available at https://github.com/caiyiqing/IF-CKT.

Introduction

Crowd counting aims to estimate the number of pedestrians in an image, which is one of the most important tasks in computer vision. For example, during the current COVID-19 pandemic, it can play a substantial role in monitoring crowd gathering and reducing the spread of the virus. Taking advantage of the deep learning techniques, previous crowd counting methods based on convolutional neural networks (CNNs) have achieved outstanding performance (Zhang et al. 2019; Qiu et al. 2019; Song et al. 2021; Lin et al. 2022). However, these fully-supervised models have been criticized for the lack of generalization ability, which suffer from severe performance degradation when being deployed into the wild. This is due to the bias towards the data distribution of the training domain. Moreover, collecting sufficient annotations for each new domain are labor-intensive, which restricts its real applications.

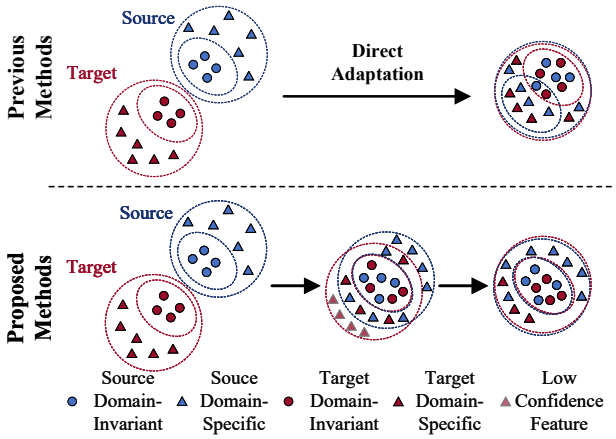

Figure 1: Comparison of previous cross-domain crowd counting methods and our proposed method. Top: Previous approaches attempt to align directly with the overall distribution while the domain invariant parts do not match. Bottom: In our method, domain-invariant features are first aligned, and then domain-invariant features serve as an intermediate bridge to further align task-related domain-specific features.

To improve the scalability of the model on the unlabeled target domain, unsupervised domain adaptation (UDA) methods have been proposed to transfer the domain-invariant information from the labeled source domain to the related but unlabeled target domain (Weiss, Khoshgoftaar,

*These authors contributed equally.
†Shaohui Lin, Changbo Wang and Gaoqi He are the corresponding authors.

and Wang 2016). Recently, UDA is applied for cross-domain crowd counting (CDCC), which are generally fallen into three categories, that is, style-transfer based, distribution-align based and pseudo-label based. The first category transforms the source synthetic image into a photo-realistic intermediate image which is then trained in a supervised manner (Wang et al. 2019, 2021; Gao et al. 2021). The second category aims to align the feature distributions between the source and target domains via one or more discriminators (Gao, Yuan, and Wang 2020; Han et al. 2020; Li, Yongbo, and Xiangyang 2019; Wu, Wan, and Chan 2021; Zou et al. 2021; Wang et al. 2022). The third category generates useful pseudo labels for fine-tuning purposes to retrain the counter, which compromises on the accuracy and the incurred noisy pseudo labels (Liu, Durasov, and Fua 2022; Cai et al. 2021).

Despite the dramatic performance improvement in reducing the domain discrepancy, these CDCC methods have not explicitly considered the transferability of different features between source and target domains. For example, different domains have their own unique scene attributes, including background, illuminations, and painting styles, which are called "domain-specific features". Correspondingly, we find that different domains share common crowd content, i.e., crowd structural features and distribution patterns, which are called "domain-invariant features". As shown in Figure 1, it is difficult for previous methods to achieve effective knowledge transfer in practice because the domain-invariant features of different domains may still be inconsistent even if the global similarity condition is satisfied. So feature confusion arises in both intra-domain and inter-domain, which results in many incorrect density estimates and quite large domain gap.

To address the aforementioned problem, we propose an innovative explicit Invariant Feature induced Cross-domain Knowledge Transformation (IF-CKT) framework. Main idea is to explicitly extract domain-invariant features as a bridge connecting the source and target domains. The whole framework consists of three parts: Global Feature Decoupling (GFD), Relation Exploration and Alignment (REA), and Graph-guided Knowledge Enhancement (GKE). Specifically, the domain-invariant and domain-specific features are first decoupled through the GFD module, which allows the model to distinguish crowds feature from backgrounds feature in complex scenes at the semantic-level. In the REA module, we first project these decoupled features into nodes to construct inter-domain relation graphs (Inter-RG) and intra-domain relation graphs (Intra-RG). Inter-RG aggregates multi-scale domain-invariant features between two domains as a bridge connecting source and target domains. Meanwhile, Inter-RG further aligns local-level invariant features, which promotes global-level feature decoupling. Intra-RG fully mines the relationship between specific features and invariant features, preserving task-related specific information as much as possible to assist the domain alignment. To further enhance the adaptability of the target domain, a GKE is established to model the confidence of pseudo-labels. In detail, the high-confidence pseudo-labels are fine-tuned to improve the adaptability of the target domain, while low-confidence pseudo-labels narrow the domain difference through adversarial learning. Main contributions are summarized as follows:

- An innovative IF-CKT is proposed to extract robust domain-invariant features as a bridge connecting source and target domains. To the best of our knowledge, IF-CKT is the first cross-domain crowd counting model that utilizes graphs to explicitly reason the relationships and interactions between domain-invariant and domain-specific features.
- A graph-guided knowledge enhancement strategy is proposed to effectively model the confidence of pseudo-labels in target domain, which further alleviate the intra-domain scale shift and distribution deviation.
- Extensive experiments indicate that IF-CKT achieves state-of-the-art performance over the existing mainstream methods on the standard benchmarks. Further, the performance of our model can be compared with those of fully supervised models trained on the target dataset.

Related Work

Since we solve the problem of cross-domain crowd counting, we first review recent works; our major contribution is to exploit graphs to explicitly reason about the relationships and interactions between domain-invariant and domain-specific features. Therefore, we also discuss related work in the above two areas.

Cross Domain Crowd Counting

Cross-domain crowd counting methods can be summarized into three strategies: style-transfer methods, feature-level adaptation methods and self-supervised methods. The style-transfer methods (Wang et al. 2019, 2021; Gao et al. 2021, 2019) narrow the domain gap by translating the synthetic images into photo-realistic images, but it is limited by the performance of the translation method. The feature-level adaptation methods (Gao, Yuan, and Wang 2020; Han et al. 2020; Li, Yongbo, and Xiangyang 2019; Wu, Wan, and Chan 2021; Zou et al. 2021; Wang et al. 2022) measure the domain discrepancy by one or more discriminators to make data distributions across domains closer. The self-supervised methods (Liu, Durasov, and Fua 2022; Cai et al. 2021) generate useful pseudo-labels on the target real images for fine-tuning purposes to retrain the counter. However, these methods optimize domain-invariant the domain-specific features as a whole. This will lead to misplacement of background features and crowd features in the scene. To this end, our method explicitly extract domain-invariant features as the bridge connecting the source domain and target domain to gradually achieve fine-grained knowledge transfer.

Feature Decoupling Learning in CDCC

Feature Decoupling learning (Locatello et al. 2019; Peng et al. 2019) aims to learn a decoupled representation that keeps latent variables separate and interpretable for the variations in the data. Due to its superior interpretability, feature decoupled learning has been well explored in tasks of few-shot learning (Ridgeway and Mozer 2018; Scott, Ridgeway,

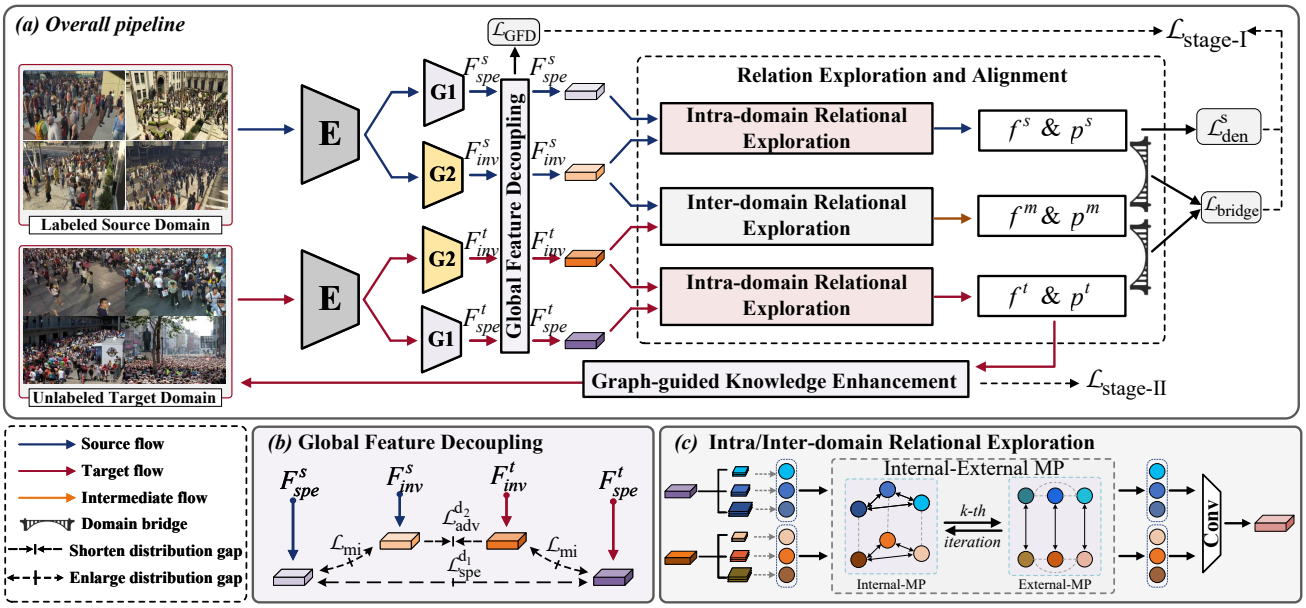

Figure 2: Illustration of our method. (a) Overall pipeline. Inputting a mini-batch of images, GFD facilitates the domain feature generators G1 and G2 to decouple the domain-specific and domain-invariant features of the two domains. In the REA module, the decoupled features are first projected into nodes to construct inter-RG and intra-RG to mine the relationship between domain-invariant features and domain-specific features. Through the $\mathcal{L}_{\text{bridge}}$, task-related domain-specific features are further aligned. Finally, we further enhance the adaptability of the target domain based on the GKE strategy. (b) Detailed structure of the GFD module. (c) Relationship exploration of REA module. We capture different topological relationships through an internal-external message passing (MP) scheme.

and Mozer 2018) , image translation (Lee et al. 2018), and object detection (Wu et al. 2022a; Liu et al. 2022). Recently, (Cheng et al. 2021; Han et al. 2022) explored the concept of feature decoupling learning in crowd counting and achieved decent performance. As for cross-domain crowd counting, on the one hand, we should effectively align domain-invariant features, on the other hand, we should be aware of the significance of task-related domain-specific features in CDCC. For example, the specific colorization information can assist the model to distinguish crowd and background in crowd counting (Bai, Wen, and Chan 2021). Therefore, the task of applying decoupling learning to CDCC is nontrivial. In this paper, we propose a novel decoupling and alignment framework that aligns invariant features while further preserving task-related domain-specific information.

Graph-Based Neural Networks in CDCC

The core parts of graph neural network constitutes nodes, edges, and the parametric information transmission functions. Graph Neural Network has been leveraged to update and reinforce the node representation by propagating information between neighbors. Because of the effectiveness and interpretability of GCN, it has been used for modeling the structural relationships of various input data, such as object detection (Zhai et al. 2021; Chen et al. 2021), action recognition (Yan, Xiong, and Lin 2018), video captioning (Yan et al. 2022), human Re-ID (Yan et al. 2019), scene understanding (Xu et al. 2017), image segmentation (Xie et al. 2021), *etc*. Some methods have attempted to leverage GNN to realize supervised crowd counting. Luo *et al.* (Luo et al. 2020) implemented GNN by interweaving multi-scale features of crowd density with its auxiliary task. Meng *et al.* (Meng et al. 2022) uses GCN to reason about the relationships between spatially-aware density features with similar density levels. Wu *et al.* (Wu et al. 2022b) proposed a spatial-temporal graph network to learn pixel-level and patch-level relationship between different domains. Although these methods have achieved promising results, their graph networks are restricted to supervised crowd counting. In contrast to (Luo et al. 2020; Meng et al. 2022; Wu et al. 2022b), our IF-CKT is considerably different, with its own advantages: 1) IF-CKT applies GCN to unsupervised domain-adaptive crowd counting for the first time. 2) Our approach fully explores the topological relationships between the inter-domain and intra-domain.

Proposed Method

The framework of IF-CKT is illustrated schematically in Figure 2, which mainly consists of three components: Global Feature Decoupling, Relation Exploration and Alignment and Graph-guided Knowledge Enhancement. In this section, we introduce the proposed framework and key techniques in detail.

Problem Formulation

Denote a labeled source domain with N_s images as $X_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, where y_i^s represents the corresponding real-valued density map in each image x_i^s. Then, an unlabeled target domain with N_t images is defined as $X_t = \{(x_i^t)\}_{i=1}^{N_t}$ with a different data distribution from the X_s. Our IF-CKT aims at transferring the knowledge from the labeled X_s to the unlabeled X_t and achieving competitive counting performance on the target domain.

Global Feature Decoupling

To facilitate the model to separate domain-invariant features (crowd structural features and distribution patterns) and domain-specific features (background features and scene style) in complex scenes, we propose to decouple global features through a GFD module. Detailed illustrations of the GFD module are shown in Figure 2 (b). Concretely, given an input image x^s and x^t, we first obtain a global domain feature map that is the output of a basic feature encoder E. Then, two different extractors $G1, G2$ are devised to decouple the domain-invariant and domain-specific features from the domain features. The processes are shown as follows:

$$\begin{aligned} F_{inv}^s &= G_2(E(x^s)), \quad F_{spe}^s = G_1(E(x^s)), \\ F_{inv}^t &= G_2(E(x^t)), \quad F_{spe}^t = G_1(E(x^t)), \end{aligned} \tag{1}$$

here, F_{inv}^s, F_{inv}^t separately indicate the domain-invariant features across the source and target domains, and F_{spe}^s, F_{spe}^t represent the domain-specific features with respect to the source and target domains. To further enhance the orthogonality within each domain, we propose a joint optimization strategy, which aligns the distributions of the domain-invariant features between the source and target domains, and enlarges the discrepancy between the domain-invariant and domain-specific features within each domain. Specifically, we first define two domain discriminators D_1 and D_2, which take F_{spe} and F_{inv} as the input, respectively. Then the discriminator outputs a domain label that indicates the source domain (0) or target domain (1). During training, the domain-specific classification loss $\mathcal{L}_{\text{spe}}^{\text{d1}}$ and domain-invariant adversarial loss $\mathcal{L}_{\text{adv}}^{\text{d}_2}$ are define as:

$$\mathcal{L}_{\text{spe}}^{\text{d}_1} = \min_{\theta_D, \theta_{G_1}} \mathcal{L}_{\text{ce}}(D_1(F_{spe}^s), 0) + \mathcal{L}_{\text{ce}}(D_1(F_{spe}^t), 1), \tag{2}$$

$$\mathcal{L}_{\text{adv}}^{\text{d}_2} = E\left[\log\left(D_2\left(F_{inv}^s\right)\right) + \log\left(1 - D_2\left(F_{inv}^t\right)\right)\right], \tag{3}$$

where $\mathcal{L}_{\text{ce}}$ measures the cross entropy between the feature maps and labels, while $\theta_{G_1}, \theta_{G_2}$ indicates the parameters for the feature extractor G_1, G_2. Inspired by (Belghazi et al. 2018), we devise a mutual information neural estimator to compute $\mathcal{L}_{\text{mi}}$. By minimizing the $\mathcal{L}_{\text{mi}}$, the orthogonality between F_{spe}^s and F_{inv}^s will be enhanced, which could promote F_{inv}^s and F_{inv}^t to contain more domain-invariant information. The loss function for the GFD module is defined as:

$$\mathcal{L}_{\text{GFD}} = \mathcal{L}_{\text{spe}}^{\text{d}_1} + \mathcal{L}_{\text{adv}}^{\text{d}_2} + \mathcal{L}_{\text{mi}}. \tag{4}$$

Relation Exploration and Alignment

In this subsection, we construct two Intra-RGs and one Inter-RG based on decoupled domain features. Inter-RG further aligns local-level invariant features, which can assist global-level feature decoupling. Meanwhile, Intra-RG fully mines the relationship between specific features and invariant features, preserving task-related specific information as much as possible. The details are described as follows:

Step 1: Graph Construction. Given the decoupled visual features $F_{inv}^s, F_{spe}^s, F_{inv}^t$ and F_{spe}^t extracted from GFD module, we perform the Pyramid Pooling Module (PPM) to extract multi-scale visual features and then utilize an interpolation layer to ensure the multi-scale feature maps to have the same size $H \times W$. Then, we aim to project the multi-scale visual features to node domain, i.e., constructing intra-source relation graph $\mathcal{G}^{IS} = \{\mathcal{V}_s^1, \mathcal{V}_s^2, \mathcal{E}\}$, intra-target relation graph $\mathcal{G}^{IT} = \{\mathcal{V}_t^1, \mathcal{V}_t^2, \mathcal{E}\}$, and inter-domain relation graph $\mathcal{G}^{ID} = \{\mathcal{V}_s^1, \mathcal{V}_t^2, \mathcal{E}\}$. $\mathcal{V}_s^i$ and $\mathcal{V}_t^i$ denotes the domain-specific and domain-invariant node feature of the source and target domain, respectively, where $i \in \{1, 2\}$. Accordingly, there are two types of edges $\mathcal{E} = \mathcal{E}_{i,j} \cup \bar{\mathcal{E}}_{m,n}$, $\mathcal{E}_{i,j} = \left\{e_{i,j}^m = \left(v_i^m, v_j^m\right)\right\}_{i,j=1}^N$ connects i_{th} scale nodes to the j_{th} scale nodes within the same domain $m \in \{1, 2\}$. $\bar{\mathcal{E}}_{m,n} = \{\bar{e}_i^{m,n} = (v_i^m, v_i^n)\}_{i=1}^N$ links same scale nodes between two domains, where $m, n \in \{1, 2\}$ and $m \neq n$.

These three graphs are treated identically for the feature update rules. Therefore, for a better readability, we will take the inter-domain relation graph $\mathcal{G}^{ID}$ as an example.

Step 2: Initial Node and Edge States. In $\mathcal{G}^{ID}$, each node is projected to the initial node embedding, namely v_i^1 and v_i^2, where $v_i^1 \in \mathcal{V}_s^1$ and $v_i^2 \in \mathcal{V}_t^2$. First, each node representation is fed into a fully connected network to compute outgoing messages, that is:

$$h_i^{1(0)} = \phi_{send}^1(v_i^1), \quad h_i^{2(0)} = \phi_{send}^2(v_i^2), \tag{5}$$

where ϕ_{send}^i is a trainable send head that has shared weights across nodes of each type. After that, the following function is used to define the edge embedding between nodes:

$$\begin{aligned} e_{i,j}^{m(k)} &= Sig(F_{edge}^1(h_i^{m(k)} - h_j^{m(k)})), \\ \bar{e}_i^{m,n(k)} &= Sig(F_{edge}^1(F_{edge}^2(h_i^{m(k)} \| h_i^{n(k)}) - h_i^{n(k)})), \end{aligned} \tag{6}$$

where F_{edge}^1 and F_{edge}^2 represent the convolution operation that is used to learn the edge embedding. "$\|$" denotes a concatenation and $Sig(\cdot)$ is the sigmoid function which maps the edge embedding into the weight value.

Step 3: Node and Edge Updating by Internal-External Message Passing. Given the initialized $\mathcal{G}^{ID}$, two different message aggregation schemes are used to compute the aggregated incoming messages $\hat{\mathbf{m}}_i^{m(k)}$ and $\bar{\mathbf{m}}_i^{m(k)}$. For the edge $e_{i,j}^m$ passed between same domain, we have the following internal message passing:

$$\hat{\mathbf{m}}_i^{m(k)} = \phi_{\text{rec}}^1\left(\sum_{(i,j)\in e_{i,j}^m} \boldsymbol{W}_e^1 e_{i,j}^{m(k-1)} h_j^{m(k-1)}\right), \tag{7}$$

where ϕ_{rec}^1 is a trainable receive head, and $\boldsymbol{W}_e^1$ denotes linear transformation matrices.

To reduce the significant discrepancy between different domains in feature space, $\bar{\mathbf{m}}_i^{m(k)}$ is computed by the following external message passing:

$$\begin{aligned}\bar{\mathbf{m}}_i^{m(k)} &= \phi_{\text{rec}}^2\left(\sum_{i=1}^{N} \bar{e}_i^{m,n(k-1)} \times \boldsymbol{\Lambda}\right), \text{where} \\ \boldsymbol{\Lambda} &= \left(\boldsymbol{W}_e^2(h_i^{m(k-1)} \| h_i^{n(k-1)})\right),\end{aligned} \tag{8}$$

where ϕ_{rec}^2 is a trainable receive head for cross-domain edges, and $\boldsymbol{W}_e^2$ denotes linear transformation matrices.

After the overall incoming message $\mathbf{m}_i^{m(k)} = \hat{\mathbf{m}}_i^{m(k)} + \bar{\mathbf{m}}_i^{m(k)}$ is got, Gated Recurrent Unit (GRU) (Li et al. 2016) is applied as the update function to obtain the final hidden features $h_i^{m(k)}$:

$$h_i^{m(k)} = GRU(h_i^{m(k-1)}, \mathbf{m}_i^{m(k)}). \tag{9}$$

After all nodes are updated, the edge weight $e_{i,j}^{(k+1)}$ for the $(k+1)_{th}$ layer is obtained by calculating pairwise contextual coefficient between nodes i and j:

$$e_{i,j}^{(k+1)} = \frac{\exp\left((h_j^{(k+1)})^{\mathrm{T}} h_i^{(k+1)}\right)}{\sum_{(i,q)\in\mathcal{E}} \exp\left((h_q^{(k+1)})^{\mathrm{T}} h_i^{(k+1)}\right)}. \tag{10}$$

Step 4: Feature Readout. After K message passing iterations, the updated multi-scale features of two node sets $H_1 = \left\{h_i^1\right\}_{i=1}^{|V^1|}$ and $H_2 = \left\{h_i^2\right\}_{i=1}^{|V^2|}$ are merged to form final feature predictions f^m of $\mathcal{G}^{IT}$:

$$f^m = M_3(cat([M_1(cat\,(H_1)), M_2(cat\,(H_2))])), \tag{11}$$

where M_i are the mapping function and $i \in \{1, 2, 3\}$. cat is the merge function by concatenation. Therefore, we get the output feature map of three graphs $\mathcal{G}^{IS}$, $\mathcal{G}^{IT}$ and $\mathcal{G}^{ID}$, including f^s, f^m and f^t, respectively.

Finally, f^s, f^m and f^t are fed to a decoder $\mathcal{D}$ to get the final density map p:

$$p^k = \mathcal{D}(f^k), k \in \{s, m, t\}. \tag{12}$$

Step 5: Invariant Feature Induced Domain Alignment. Through relational exploration in the above steps, Intra-RG and inter-RG fully mine the relationship between specific features and invariant features. To further align task-related domain-specific information, we implement a Invariant induced bridging loss at both feature-level and density-level. For the feature level, we use the L2-norm loss to measure the features' distance:

$$\mathcal{L}_f = \sum_{k\in\{s,t\}} \left\|f^k - f^m\right\|^2. \tag{13}$$

For the density level, we use the cross-entropy to measure the distribution gap:

$$\mathcal{L}_p = \sum_{k\in\{s,t\}} E\left[\log\left(D_{\text{inv}}\left(p^k\right)\right) + \log\left(1 - D_{\text{inv}}\left(p^m\right)\right)\right], \tag{14}$$

where D_{inv} denotes the discriminator. The final bridge losses can be formulated as:

$$\mathcal{L}_{\text{bridge}} = \gamma_1\mathcal{L}_p + \gamma_2\mathcal{L}_f, \tag{15}$$

where γ_1 and γ_2 are the weights to balance the two losses.

Graph-Guided Knowledge Enhancement

The target domain data collected from the real world usually has extreme data distribution. These distributions are caused by various factors such as moving objects, weather conditions and camera parameters. These extreme distributions lead to a large gap between domain invariant features and domain specific features in the target domain. Therefore, we design GKE to generate high confidence pseudo labels for the target domain, so as to further improve the adaptability of domain-invariant features in the target domain.

We take advantage of the edge embedding to determine the confidence of the target density map. A novel ranking method is designed by using the following function:

$$R(x^t) = \frac{1}{N \times N + N}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} e_{i,j}^m + \sum_{i=1}^{N} \bar{e}_i^{m,n}\right), \tag{16}$$

which is the mean value of edge embedding value. Given a ranking of scores from $R(x^t)$, hyperparameter λ is introduced as a ratio to separate the target images into a low-confidence and a high-confidence split. Let x^{te} and x^{th} denote a target image assigned to the high-confidence and low-confidence split, respectively. In order to conduct domain separation, we define $R(x^{te}) < \lambda$ and $R(x^{th}) >= \lambda$. For high-confidence, we use their complete pseudo labels:

$$\mathcal{L}_{den}^p = \left\|p_h^t - \hat{p}_h^t\right\|, \tag{17}$$

where $\hat{p}_h^t$ is the pseudo label of the target domain generated in the previous stage, and p_h^t is the density map of the high-confidence samples from target domain.

For the low-confidence, adversarial learning is used to enforce the feature alignment:

$$\mathcal{L}_{\text{adv}}^p = E\left[\log\left(D_{spe}\left(p_l^t\right)\right) + \log\left(1 - D_{spe}\left(\hat{p}_l^t\right)\right)\right], \tag{18}$$

where D_{spe} denotes the discriminator. p_l^t is the density map of the low-confidence samples from target domain and $\hat{p}_l^t$ is the corresponding pseudo label.

Final Objective Function Optimization

We integrate the losses as mentioned above. The total loss of the stage-I (IF-CKT) could be formulated as:

$$\mathcal{L}_{\text{stage-1}} = \mathcal{L}_{GFD} + \mathcal{L}_{\text{den}}^s + \mathcal{L}_{\text{bridge}}, \tag{19}$$

where $L_{\text{den}}^s = \left\|f^s - y^s\right\|^2$. The total loss of the stage-II (GKE) could be formulated as:

$$\mathcal{L}_{\text{stage-2}} = \mathcal{L}_{\text{den}}^p + \gamma_3\mathcal{L}_{adv}^p. \tag{20}$$

Finally, our complete loss function is formed by all the loss functions:

$$\mathcal{L} = \mathcal{L}_{\text{stage-1}} + \mathcal{L}_{\text{stage-2}}. \tag{21}$$

Our model has two optimization stages. In the stage-I, we train optimized GFD and REA modules. In stage-II, we leverage the GKE strategy to further enhance the adaptability of the target domain. γ_1, γ_2 and γ_3 were set to 1, 0.3 and 1, respectively, by cross-validation.

Method	TS	SHT B		WorldExpo'10 (MAE)						UCF-QNRF		MALL	
		MAE	MSE	S1	S2	S3	S4	S5	Avg.	MAE	MSE	MAE	MSE
MCNN (Zhang et al. 2016b)	yes	26.4	41.3	3.4	20.6	12.9	13.0	8.1	11.6	277	426	2.24	8.5
IG-CNN (Sam et al. 2018)	yes	13.6	21.1	2.6	16.1	10.15	20.2	7.6	11.3	-	-	-	-
CycleGAN (Zhu et al. 2017)	no	25.4	39.7	4.4	69.6	49.9	29.2	9.0	32.4	257.3	400.6	-	-
SE CycleGAN (Wang et al. 2019)	no	19.9	28.3	4.3	59.1	43.7	17.0	7.6	26.3	230.4	384.5	-	-
FA (Gao, Yuan, and Wang 2020)	no	16.0	24.7	5.7	59.9	19.7	14.5	8.1	21.6	-	-	-	-
IDK (Cai et al. 2021)	no	14.3	22.8	-	-	-	-	-	-	224.3	375.8	-	-
FSC (Han et al. 2020)	no	16.9	24.7	4.2	54.7	40.5	**10.5**	36.4	29.3	221.2	390.2	2.47	3.25
IFS (Gao et al. 2019)	no	13.1	19.4	4.5	33.6	14.1	30.4	4.4	17.4	211.7	357.9	2.31	2.96
DACC (Gao et al. 2021)	no	13.1	19.4	4.5	33.6	14.1	30.4	4.4	17.4	203.5	343.0	2.31	2.96
BLA (Gong et al. 2022)	no	11.9	18.9	-	-	-	-	**-**	17.9	198.9	316.1	-	-
CDCC (Liu, Durasov, and Fua 2022)	no	11.4	17.3	4.0	31.9	23.5	19.4	**4.2**	16.6	198.3	332.9	-	-
NoAdpt	no	22.4	30.2	5.4	82.2	62.1	22.2	14.3	30.5	275.4	450.3	4.27	5.35
IF-CKT	no	12.3	18.4	4.4	38.3	17.4	21.1	13.5	18.9	194.5	330.3	2.45	3.20
IF-CKT+	no	**10.9**	**16.8**	**4.1**	**32.5**	**13.1**	20.0	12.4	**16.4**	**190.6**	**324.9**	**2.14**	**2.71**

Table 1: The performance of other domain adaptation methods and our method on the four real-world datasets. (TS:Target Supervision)

Experiments

To demonstrate the superiority of our method, we conduct extensive experiments on four datasets, including ShanghaiTech B dataset (Zhang et al. 2016b), WorldExpo'10 dataset (Zhang et al. 2016a), UCF-QNRF dataset (Idrees et al. 2018) and MALL dataset (Chen et al. 2012), Following (Zhang et al. 2016b), we adopt Mean Absolute Error (MAE) and Mean Squared Error (MSE) as the evaluation metrics.

Implementation Details

For fair comparisons with previous methods, we chose the first 13 layers from the VGG-16 (Simonyan and Zisserman 2015) network as the basic feature encoder E. For the domain discriminators, we respectively design a network which includes five convolutional layers with stride of 1 and kernel size 3, the channels of each layer are 512, 256, 128, 64, 1 respectively. For the MI estimators, we separately utilize a network consisting of three fully-connected layers. The G is trained using the Stochastic Gradient Descent (SGD) optimizer with a learning rate as 10^{-6} . We use Adam optimizer (Kingma and Ba 2015) with learning rate of 10^{-4} for the discriminators. For data generation and augmentation, we follow the commonly used methods introduced in MCNN (Zhang et al. 2016b).

Comparisons with State-of-the-Art Method

We compare our method with the previously published cross-domain crowd counting methods under the adaptation scenarios from synthetic GCC dataset to four different real-world datasets. For each pair of datasets, we report the errors between the generated density maps and the ground truth maps on the target set. Several variants of the proposed model are defined: 1) NoAdapt: the model is only trained on the source samples. 2) IF-CKT: Only perform the distribution alignment in the first stage. 3) IF-CKT+: the full model with the GKE strategy. As shown in Table 1, although the complexity of the scenarios for these real-world datasets, we performed best on all the target domains. Our IF-CKT+ reduced counting errors by 0.5, 0.2, and 7.7 in MAE, compared to the previous best DA results (CDCC) on three real-world datasets (SHT B, WorldExpo'10 and UCF- QNRF). **Quantitative evaluation of the learned decoupling features are given in the supplementary material.**

NA	IF-CKT			GKE		MAE	MSE
	GFD	RE	L_{bri}	w/a	w/r		
✓						275.4	450.3
	✓					225.3	381.5
	✓	✓				209.7	354.4
	✓	✓	✓			200.4	340.8
	✓	✓	✓	✓		194.5	330.3
	✓	✓	✓	✓	✓	190.6	324.9

Table 2: Effects of different model components in GCC to UCF-QNRF setting. {GFD, RE, L_{bri}} correspond to {GFD module, relation exploration, bridge loss} respectively. w/a means use all pseudo labels to retrain, and w/r means use our ranking method where r=0.74. NA means no adapt methods.

Qualitative results of the estimated density maps can be seen in Figure 3. Due to the significant differences among different domains, NoAdapt can only reflect the crowd distribution trend while failing to align domain-specific features and domain-invariant features. Differently, IF-CKT can consistently estimate more accurate crowd density due to the proposed IF-CKT framework. It is obvious that IF-CKT+ vastly promote the quality of the predicted density maps. **Due to limited space, qualitative analysis on the disentanglement effectiveness of our full method are given in the supplementary material.**

Ablation Study

Analysis of Different Components We analyzed the effect of each component in the proposed method. As listed in Table 2, each module was eliminated to verify the utilization effectiveness of the all-combined modules. It is

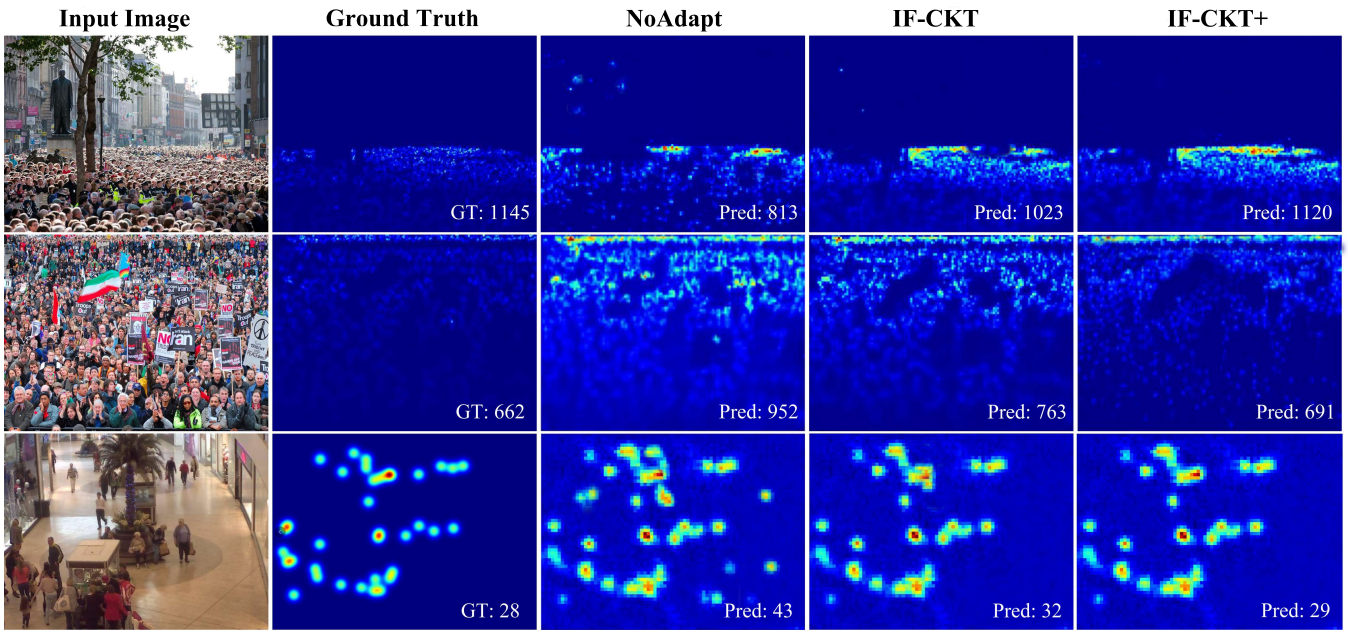

Figure 3: Visualization performance of experiment. Row 1 and 2 come from Shanghai Tech, and row 3 was from MALL.

λ	0	0.5	0.6	0.7	**0.74**	0.8
MAE	198.7	194.9	193.4	191.8	**190.6**	192.4
MSE	339.2	331.2	329.8	326.5	**324.9**	327.6

Table 3: Analysis of Hyperparameter λ in UCF-QNRF dataset.

shown that the final performance was gradually improved with the addition of each component. Specifically, as for the "IF-CKT" model, the estimation errors reduced remarkably, whether MAE (from 275.4 to 200.4) or MSE (from 450.3 to 340.8). "GKE" strangey improved the performance from 200.4/340.8 to 190.6/324.9.

Analysis of Hyperparameter λ We experimented on determining a proper value for the hyperparameter λ. In Table 3, different values of λ are used for target domain separation. When $\lambda = 0.74$, the model achieves the best performance (190.6 on MAE and 324.9 on MSE) under GCC to QNRF setting.

Analysis of Domain Hyperparameter N and K As listed in Table 4, the performance of the model improves significantly (MAE: 12.3 to 10.9 and MSE: 18.5 to 16.8) when the model applies more nodes (2 to 3). In addition, the model shows limited improvement when considering more nodes (3 to 5). Furthermore, Table 4 illustrates the performance of our model trained with different iterations K. The performance of our final model peaks at training with three iterations, and the performance gradually degrades later. This is because as the iterations K increase, noisy messages start to permeate through the graph and hamper the final prediction. Therefore, we choose $N = 3$ and $K = 3$.

Domain Graph Setting	MAE	MSE
(N=2,K=3)	12.3	18.5
(N=3,K=3)	10.9	16.8
(N=5,K=3)	10.9	16.7
(N=3,K=1)	13.1	19.6
(N=3,K=3)	10.9	16.8
(N=3,K=5)	10.9	16.9
IF-CKT	**10.9**	**16.8**

Table 4: Analysis of the domain graph hyperparameter N and K. Results are obtained under GCC to SHT B setting.

Conclusion

In this study, a novel IF-CKT framework is proposed for the CDCC problem by exploring cross-domain topological relationships. Domain-invariant features of different domains are explicitly modeled and worked as the efficient bridge connecting the two domains. Inter-RG and Intra-RG are designed to extract domain-invariant features while further utilizing task-related domain-specific information to assist domain alignment. Extensive migration experiments indicate that IF-CKT achieved state-of-the-art performance over the existing mainstream methods on standard benchmarks. To further verify the effectiveness of our method, we also evaluate the cross-dataset experimental results and discuss the complexity of the model.

In future work, we will further consider crowd features at different density levels to achieve more fine-grained feature decoupling and relational reasoning in CDCC.

Acknowledgments

This work was supported in part by Natural Science Foundation of Chongqing (No. CSTB2022NSCQ-MSX0552), National Natural Science Foundation of China (No. 62002121, 62072183, and 62102151), Shanghai Science and Technology Commission (No. 21511100700, 22511104600), the National Key Research and Development Program of China (No. 2021ZD0111000), the Research Project of Shanghai Science and Technology Commission (No. 20DZ2260300), Shanghai Sailing Program (21YF1411200) and CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2021-031A), the Open Project Program of the State Key Lab of CAD&CG (No. A2203), Zhejiang University.

References

Bai, H.; Wen, S.; and Chan, S.-H. G. 2021. Crowd Counting by Self-supervised Transfer Colorization Learning and Global Prior Classification. *arXiv preprint arXiv:2105.09684*.

Belghazi, M. I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, R. D. 2018. Mine: mutual information neural estimation. In *international conference on machine learning*.

Cai, Y.; Chen, L.; Ma, Z.; Lu, C.; Wang, C.; and He, G. 2021. Leveraging Intra-Domain Knowledge to Strengthen Cross-Domain Crowd Counting. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Chen, C.; Li, J.; Zheng, Z.; Huang, Y.; Ding, X.; and Yu, Y. 2021. Dual Bipartite Graph Learning: A General Approach for Domain Adaptive Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2703–2712.

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

Cheng, J.; Xiong, H.; Cao, Z.; and Lu, H. 2021. Decoupled two-stage crowd counting and beyond. *IEEE Transactions on Image Processing*, 30: 2862–2875.

Gao, J.; Han, T.; Wang, Q.; and Yuan, Y. 2019. Domain-adaptive crowd counting via inter-domain features segregation and gaussian-prior reconstruction. *arXiv preprint arXiv:1912.03677*.

Gao, J.; Han, T.; Yuan, Y.; and Wang, Q. 2021. Domain-Adaptive Crowd Counting via High-Quality Image Translation and Density Reconstruction. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.

Gao, J.; Yuan, Y.; and Wang, Q. 2020. Feature-aware adaptation and density alignment for crowd counting in video surveillance. *IEEE transactions on cybernetics*, 51(10): 4822–4833.

Gong, S.; Zhang, S.; Yang, J.; Dai, D.; and Schiele, B. 2022. Bi-level Alignment for Cross-Domain Crowd Counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7542–7550.

Han, T.; Bai, L.; Gao, J.; Wang, Q.; and Ouyang, W. 2022. DR. VIC: Decomposition and Reasoning for Video Individual Counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3083–3092.

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

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.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; and Yang, M.-H. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, 35–51.

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

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2016. Gated graph sequence neural networks. In *International Conference on Learning Representations*.

Lin, H.; Ma, Z.; Ji, R.; Wang, Y.; and Hong, X. 2022. Boosting Crowd Counting via Multifaceted Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19628–19637.

Liu, D.; Zhang, C.; Song, Y.; Huang, H.; Wang, C.; Barnett, M.; and Cai, W. 2022. Decompose to Adapt: Cross-domain Object Detection via Feature Disentanglement. *IEEE Transactions on Multimedia*, 1–1.

Liu, W.; Durasov, N.; and Fua, P. 2022. Leveraging Self-Supervision for Cross-Domain Crowd Counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5341–5352.

Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, 4114–4124. PMLR.

Luo, A.; Yang, F.; Li, X.; Nie, D.; Jiao, Z.; Zhou, S.; and Cheng, H. 2020. Hybrid graph neural networks for crowd counting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 11693–11700.

Meng, Y.; Bridge, J.; Wei, M.; Zhao, Y.; Qiao, Y.; Yang, X.; Huang, X.; and Zheng, Y. 2022. Counting with Adaptive Auxiliary Learning. *arXiv preprint arXiv:2203.04061*.

Peng, X.; Huang, Z.; Sun, X.; and Saenko, K. 2019. Domain agnostic learning with disentangled representations. In *International Conference on Machine Learning*, 5102–5112. PMLR.

Qiu, Z.; Liu, L.; Li, G.; Wang, Q.; Xiao, N.; and Lin, L. 2019. Crowd counting via multi-view scale aggregation networks. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 1498–1503. IEEE.

Ridgeway, K.; and Mozer, M. C. 2018. Learning deep disentangled embeddings with the f-statistic loss. *Advances in neural information processing systems*, 31.

Sam, D. B.; Sajjan, N. N.; Babu, R. V.; and Srinivasan, M. 2018. 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*, 3618–3626.

Scott, T.; Ridgeway, K.; and Mozer, M. C. 2018. Adapted deep embeddings: A synthesis of methods for k-shot inductive transfer learning. *Advances in Neural Information Processing Systems*, 31.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Song, Q.; Wang, C.; Jiang, Z.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; and Wu, Y. 2021. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3365–3374.

Wang, Q.; Gao, J.; Lin, W.; and Yuan, Y. 2019. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8198–8207.

Wang, Q.; Gao, J.; Lin, W.; and Yuan, Y. 2021. Pixel-wise crowd understanding via synthetic data. *International Journal of Computer Vision*, 129(1): 225–245.

Wang, Q.; Han, T.; Gao, J.; and Yuan, Y. 2022. Neuron Linear Transformation: Modeling the Domain Shift for Crowd Counting. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8): 3238–3250.

Weiss, K.; Khoshgoftaar, T. M.; and Wang, D. 2016. A survey of transfer learning. *Journal of Big data*, 3(1): 1–40.

Wu, A.; Han, Y.; Zhu, L.; and Yang, Y. 2022a. Instance-Invariant Domain Adaptive Object Detection Via Progressive Disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4178–4193.

Wu, Q.; Wan, J.; and Chan, A. B. 2021. Dynamic Momentum Adaptation for Zero-Shot Cross-Domain Crowd Counting. In *Proceedings of the 29th ACM International Conference on Multimedia*, 658–666.

Wu, Z.; Zhang, X.; Tian, G.; Wang, Y.; and Huang, Q. 2022b. Spatial-Temporal Graph Network for Video Crowd Counting. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Xie, G.-S.; Liu, J.; Xiong, H.; and Shao, L. 2021. Scale-aware graph neural network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5475–5484.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5410–5419.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*.

Yan, Y.; Zhang, Q.; Ni, B.; Zhang, W.; Xu, M.; and Yang, X. 2019. Learning context graph for person search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2158–2167.

Yan, Y.; Zhuang, N.; Ni, B.; Zhang, J.; Xu, M.; Zhang, Q.; Zhang, Z.; Cheng, S.; Tian, Q.; Xu, Y.; Yang, X.; and Zhang, W. 2022. Fine-Grained Video Captioning via Graph-based Multi-Granularity Interaction Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2): 666–683.

Zhai, Q.; Li, X.; Yang, F.; Chen, C.; Cheng, H.; and Fan, D.-P. 2021. Mutual graph learning for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12997–13007.

Zhang, C.; Kang, K.; Li, H.; Wang, X.; Xie, R.; and Yang, X. 2016a. Data-driven crowd understanding: A baseline for a large-scale crowd dataset. *IEEE Transactions on Multimedia*, 18(6): 1048–1061.

Zhang, Y.; Zhou, C.; Chang, F.; and Kot, A. C. 2019. A scale adaptive network for crowd counting. *Neurocomputing*, 362: 139–146.

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

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zou, Z.; Qu, X.; Zhou, P.; Xu, S.; Ye, X.; Wu, W.; and Ye, J. 2021. Coarse to fine: Domain adaptive crowd counting via adversarial scoring network. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2185–2194.