

Joint CNN and Transformer Network via weakly supervised Learning for efficient crowd counting

Fusen Wang, Kai Liu, Fei Long, Nong Sang, *Member, IEEE*, Xiaofeng Xia, Jun Sang

arXiv:2203.06388v1 [cs.CV] 12 Mar 2022

Abstract—Currently, for crowd counting, the fully supervised methods via density map estimation are the mainstream research directions. However, such methods need location-level annotation of persons in an image, which is time-consuming and laborious. Therefore, the weakly supervised method just relying upon the count-level annotation is urgently needed. Since CNN is not suitable for modeling the global context and the interactions between image patches, crowd counting with weakly supervised learning via CNN generally can not show good performance. The weakly supervised model via Transformer was sequentially proposed to model the global context and learn contrast features. However, the transformer directly partitions the crowd images into a series of tokens, which may not be a good choice due to each pedestrian being an independent individual, and the parameter number of the network is very large. Hence, we propose a Joint CNN and Transformer Network (JCTNet) via weakly supervised learning for crowd counting in this paper. JCTNet consists of three parts: CNN feature extraction module (CFM), Transformer feature extraction module (TFM), and counting regression module (CRM). In particular, the CFM extracts crowd semantic information features, then sends their patch partitions to TRM for modeling global context, and CRM is used to predict the number of people. Extensive experiments and visualizations demonstrate that JCTNet can effectively focus on the crowd regions and obtain superior weakly supervised counting performance on five mainstream datasets. The number of parameters of the model can be reduced by about 67%~73% compared with the pure Transformer works. We also tried to explain the phenomenon that a model constrained only by count-level annotations can still focus on the crowd regions. We believe our work can promote further research in this field.

Index Terms—Crowd counting, CNN, Transformer, Weakly supervised learning, Count-level annotations.

I. INTRODUCTION

Crowd Counting based on deep learning is a hot topic due to its significant role in many applications, such as crowd tracking, public transportation, transfer counting, etc. The purpose of crowd counting is to obtain the counting of each image. The current mainstream works [1]–[7] usually adopt the learning method of location-level supervision and have achieved significant precision promotion. However, these works require location-level annotation of each person, which has some weaknesses: (1) Collecting the location-level person annotations can become expensive and laborious in a large crowd scenario, such as, in NWPU-Crowd [8], the image of

Fusen Wang and Kai Liu and Fei Long and Xiaofeng Xia and Jun Sang are with the School of Big Data & Software Engineering, Chongqing University, Chongqing 401331, China (201924131014@cqu.edu.cn; kailiu@cqu.edu.cn; 201924132039@cqu.edu.cn; nsang@hust.edu.cn; xiaofeng@cqu.edu.cn; jsang@cqu.edu.cn). Jun Sang is the corresponding author.

Nong Sang is with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China.



Fig. 1. Patch Partition in the input image. In the middle, several person heads are divided into two parts, which is unreasonable and may lead to overestimation or underestimation.

the largest scene contains 20,033 person head annotations. (2) Due to the gap between the training and inferencing phases in crowd counting, location-level annotations are only used for the former and redundant for the latter [9]. Therefore, weakly supervised learning with counting-level annotations for crowd counting is essential. In addition, MATT [10] pointed out that the total count of objects usually can be economically obtained in many realistic scenes.

In recent works, the counting performance of the weakly supervised model via CNN [10], [11] is relatively poor due to convolution kernels' limited receptive field and failure in modeling the global context information, which plays a vital role in crowd counting [4], [12]. Besides, CNN can not establish the interactions between image patches, so it is incapable of learning the contrast features between crowd and background. Compared with the CNN, the Transformer can adequately learn the contrast features between different image patches due to the self-attention mechanism in image classification, segmentation [13]–[15], etc. This features can be significantly used to distinguish crowds from background areas. Hence, some researchers developed transformer architecture to effectively capture crowd global context information and model long-range dependency [9], [12], [16], which all achieved superior counting performance. However, it is unreasonable for the Transformer to directly divide the input images into a series of patches as tokens. As shown in Fig. 1, the input image is divided into two patches, and several person heads

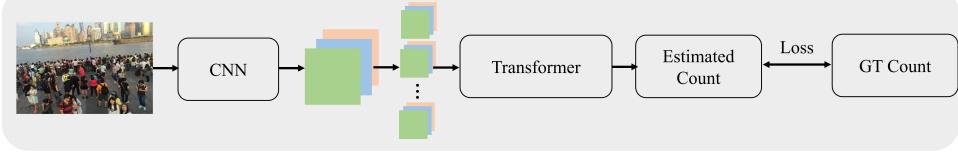


Fig. 2. The overview of the proposed Weakly supervised counting via Joint CNN and Transformer, namely JCTNet. The CNN is used to extract crowd semantic information and the Transformer is utilized to model global context and learn contrast features for improving the final counting performance.

are split into two parts, which may cause overestimation or underestimation due to each pedestrian being an independent individual. Besides, the pure transformer model needs more parameters to achieve superior performance due to the lack of CNN's inductive bias ability [13]–[15]. To sum up, CNN and Transformer may be complementary.

In this paper, we propose a Joint CNN and Transformer Network (JCTNet) for weakly supervised counting, as shown in Fig. 2, which effectively alleviates the above issues and achieves better counting performance with fewer parameters. A detailed pipeline diagram is shown in Fig. 3. JCTNet consists of three modules: CNN feature extraction module (CFM), Transformer feature extraction module (TFM), and counting regression module (CRM). Specifically, CFM utilizes VGG16-BN [17] first ten layers to extract crowd semantic features, which can relieve the issue of dividing the input images. Similar to [14], TFM includes four modified Swin Transformer blocks (MSTB) for local attention and cross-window interaction. The semantic features from CFM are transmitted into TFM to capture global context and model long-range dependency. Finally, the token features from TFM are reshaped into three dimensions and utilized to estimate the number of people by CRM.

To summarize, the main contributions of our work are outlined as follows:

- (1) We propose a Joint CNN and Transformer Network (JCTNet) with complementary learning for weakly supervised crowd counting, which only relies on count-level annotations to reduce annotation costs. In addition, this joint learning can make up for the deficiencies between CNN and Transformer for further improving the counting performance.
- (2) On five mainstream datasets (ShanghaiTech Part A/B, UCF-CC-50, UCF-QNRF, NWPU-Crowd), extensive experiments demonstrate that JCTNet obtains superior counting performance.
- (3) The proposed JCTNet has fewer parameters and achieves lower counting errors compared with the pure transformer works. As shown in Table 1, JCTNet obtains lower counting errors (MAE of 62.8) on the ShanghaiTech Part A dataset [1] and only employs 33% and 27% of the parameters of TransCrowd [9] and CCTrans [16], respectively. In addition, we analyze why this model constrained only by count-level annotations can focus on crowd regions in Section 5 and hope this work could encourage further research in the weakly supervised crowd counting.

II. RELATED WORK

In this section, we briefly review some mainstream related works on crowd counting methods via fully supervised and weakly supervised.

A. Crowd Counting methods via fully supervised

The crowd counting review [19] has introduced many excellent works. Zhang et al. proposed a Multi-Column Convolution Neural Network, composed of three branches with convolution kernels of different sizes, namely MCNN [1]. Later, Sam et al. proposed a density classifier, namely Switch CNN [20], for adaptively selecting optimal regressors from three different branches. Boominathan et al. proposed CrowdNet [2], including the shallow branch for extracting small-scale features and the deep branch for the opposite. Guo et al. employed a multi-column dilated attention network to capture different receptive fields and focus on the crowd area [3]. Later, Zhu et al. proposed a dual path fusion network, of which one path generated an attention map, and the other path generated the high-quality density map (SFANet [5]). Sindagi et al. employed a hierarchical attention-based network, including a spatial attention module and several global attention modules (HACNN [11]). Jiang et al. proposed attention scaling for Crowd Counting, namely ASNet [21], composed of density attention network (DANet) for generating attention masks of different density levels and attention scaling network (ASNet) for generating several intermediate density maps and obtaining the final density maps.

B. Crowd Counting methods via weakly supervised

Lei et al. proposed the model MATT [10] for weakly supervised crowd counting with a small amount of location-level annotations and a large amount of count-level annotations. Yang et al. [18] proposed the sorting network for directly regressing counting without location-level annotations. Sindagi et al. [11] proposed the weakly supervised model using image-level labels for adapting existing crowd counting models to new scenes and datasets. The method [22] proposed an almost unsupervised counting method for crowd counting. However, the above works based on CNN did not achieve comparable counting performance.

Recently, the transformer has achieved significant performance by modeling the global interactions between different regions in many vision tasks, such as image classification [13], [23], image segmentation [13]–[15], [24], object detection [25]. Some recent works introduce the vision transformer into weakly supervised crowd counting. Liang et al. first proposed a

TABLE I
COUNTING ERRORS OF SEVERAL WEAKLY-SUPERVISED CROWD COUNTING METHODS ON SHANGHAI TECH PART A.

Method	Publication	Model	MAE	MSE	Parameters(M)
Yang et al. [18]*	ECCV20	CNN	104.6	145.2	—
MATT [10]*	PR21	CNN	80.1	129.4	—
TransCrowd [9]*	Arxiv21	TRM	66.1	95.4	86
CCTrans [16]*	Arxiv21	TRM	64.4	95.4	104
JCTNet(ours)*	—	CNN+TRM	62.8	95.6	28

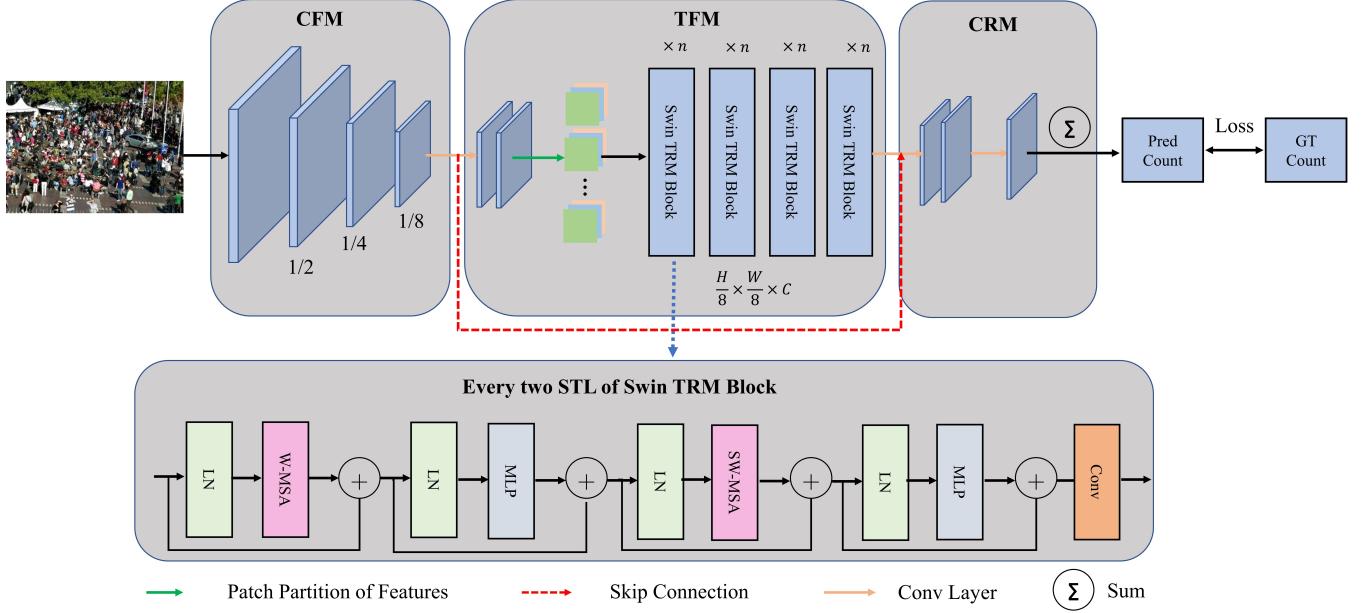


Fig. 3. The proposed JCTNet overview. First, the input image is fed into the CNN feature extraction module (CFM) to learn crowd semantic features. Then, by patch partition, these features are split into the 1D sequences, which go through the Transformer feature extraction module (TFM) for modeling global context. Finally, the outputs of TFM are reshaped into 2D feature maps, and the counting regression module (CRM) serves for predicting the counting result. STL denotes the Swin Transformer layer.

transformer-based weakly supervised crowd counting network, namely TransCrowd [9], direct regression of the number of people by the proposed TransCrowd-Token and TransCrowd-GAP. Tian et al. proposed simplifying and improving crowd counting with Transformer, namely CCTrans [16], which covered both full supervision and weak supervision method.

However, the above works usually employ the transformer as the backbone of the model. This involves too many parameters and calculations and is not conducive to future lightweight research.

III. PROPOSED APPROACHES

In this section, we first illustrate the overview architecture of our proposed JCTNet for weakly supervised counting and then introduce each component in detail. The JCTNet includes three modules: (1) CNN feature extraction module (CFM); (2) Transformer feature extraction module (TRM); (3) counting regression module, as shown in Fig. 3.

A. CNN feature extraction module

The motivation for designing CFM is as follows: (1) Some transformer works [13]–[15] directly split the input images into many patches then feed them into Transformer blocks. The process is not appropriate for the individuals in crowd

counting. (2) For weakly supervised learning, in addition, deeper crowd semantic features extracted by CNN can bring better performance than shallow features such as color, texture, etc. Hence, as shown in Fig. 3, we adopt the first ten layers of the VGG16-BN [17] for crowd semantic feature extraction and as the input of the sequent TFM.

The CFM employs VGG16 as the extractor due to its excellent performance on many computers vision tasks such as object detection [26], classification [27], crowd counting [28], etc. Besides, W-net [28] has also demonstrated that VGG-BN could generate superior performance as the backbone of crowd counting. To enlarge the receptive field while maintaining the number of parameters, we eliminate the last two pooling layers and ensure that the output of the network is 1/8 of the original image resolution. Given a crowd input image $I \in \mathbb{R}^{H \times W \times 3}$ (H , W , and 3 are respectively its height, width, and channel size), the definition of CFM is as follows:

$$C_f = \mathcal{F}_{vgg}(I) \quad (1)$$

where C_f is the feature from CNN, and \mathcal{F}_{vgg} denotes the VGG16-BN first ten layers.

B. Transformer feature extraction module

In this module, we introduce the proposed Transformer feature extraction module (TFM) using the output of CFM as the input. Now, describe the data flow of our model as illustrated in Fig. 3.

2D Feature to 1D Sequence. Obtain the crowd semantic features $C_f \in \mathbb{R}^{H \times W \times C}$ from the above CFM, and then split it into $\frac{HW}{K^2}$ image patches and the size of each patch is $K \times K \times C$. Flatten these 2D patches features into the 1D sequence token $x \in \mathbb{R}^{N \times D}$, where $N = \frac{HW}{K^2}$, $D = K \times K \times C$. We project a learnable projection $f : x_i \rightarrow e_i \in \mathbb{R}^D$ for mapping the x into embedded features, i.e., image tokens.

Modified Swin-Transformer Blocks. We adopt modified Swin-Transformer [14] as the primary transformer feature extractor due to its success in local attention and cross-window attention. The TFM contains four modified Swin Transformer blocks. In particular, we discard the patch merging layer to avoid the down-sample because the CFM has reduced the image's resolution to 1/8 of the original input. We add a convolution layer at the end of every two Swin Transformer layers of each modified Swin Transformer block to enhance the interaction between CNN and transformer features and bring the inductive bias of convolution into the transformer, motivated by SwinIR [29]. And, we add the skip connection for feature fusion between the CFM and the TFM. The output of TFM is defined by the following equation:

$$T_F = \mathcal{F}_{TFM}(Conv(C_f)) + Conv(C_f) \quad (2)$$

where \mathcal{F}_{TFM} are the Transformer feature extraction module, and T_F denotes the obtained transformer features from TFM. $Conv$ is the convolution layer at the end of the CFM for channel dimension reduction and skip connection.

The TFM consists of 4 MSTB and each MSTB includes several Swin Transformer layers (STL). The workflow of MSTB is formulated as follows:

$$\begin{aligned} T_{F_i} &= \mathcal{F}_{MSTB_i}(T_{F_{i-1}}), i = 1, 2, 3, 4 \\ T_{F_{i,j}} &= Conv(\mathcal{F}_{STL_{i,j}}(\mathcal{F}_{STL_{i,j-1}}(T_{F_{i,j-2}}))), \\ j &= 2, 4, 6, 8, \dots, L \end{aligned} \quad (3)$$

where \mathcal{F}_{MSTB_i} are the i -th MSTB, and T_{F_i} denotes the output features of each MSTB. In Eq. (4), $\mathcal{F}_{STL_{i,j}}$ is the j -th Swin Transformer layer of the i -th MSTB, and $T_{F_{i,j}}$ is the intermediate features by L Swin Transformer layers. Specially, we combine two Swin Transformer layers and one convolution layer into a whole.

Swin-Transformer layer. Based on the standard transformer module [13], Swin Transformer [14] proposed the window multi-head self-attention (W-MSA) for reducing the amount of calculation. However, the window-based self-attention lacked interaction across windows and limited its power in modeling global context. Then, Swin Transformer introduced shifted-window multi-head self-attention [14] (SW-MSA) for maintaining the efficient computation of non-overlapping windows. As shown at the bottom of Fig. 3, a standard 2-layer Swin Transformer block is the consecutive combination of W-MSA and SW-MSA, and the other layers

are kept the same. Given the input sequence $Z_{l-1} \in \mathbb{R}^{N \times D}$ of l -th layer, first, reshape it into the 2D feature $H \times W \times C$. These features can be fed into W-MSA for computing local attention and sequent be transmitted into SW-MSA to execute cross-window connections. The whole process is formulated as follows:

$$Z'_l = WMSA(LN(Z_l)) + Z_l \quad (5)$$

$$Z_l = MLP(LN(Z'_l)) + Z'_l \quad (6)$$

$$Z'_{l+1} = SWMSA(LN(Z_l)) + Z_l \quad (7)$$

$$Z_{l+1} = MLP(LN(Z'_{l+1})) + Z'_{l+1} \quad (8)$$

$$Z_{l+1} = Conv(Z_{l+1}) \quad (9)$$

where Z_l and Z_{l+1} are the output features of the consecutive W-MSA and SW-MSA transformer blocks, respectively. The MLP consists of two fully-connected layers with GELU non-linearity. And, the LN denotes the layer normalization, which is added before both MSA and MLP. The residual connection is employed for both modules. Finally, add a convolution layer for features interaction between CNN and transformer features.

C. Counting regression module

The output of the above TFM is fed to the counting regression module (CFM) to predict the final count through a set of convolution layers. These architectures are designed as follows: Conv (256,128,3,2,2)-BN-ReLU, Conv (128,64,3,2,2)-BN-ReLU, Conv (64,1,1), where Conv (*) contains input channel, output channel, kernel size, dilation rate, padding, respectively. The dilation rate is employed for enlarging receptive fields while containing the spatial resolution.

IV. IMPLEMENTATION DETAILS

In this section, we first introduce the five mainstream datasets. Sequentially, the loss function, training details are given respectively.

A. Datasets

ShanghaiTech [1] dataset contains 1,198 images with 330,165 annotated heads, divided into two parts: Part A and Part B. Part A includes 300 training images and 182 test images randomly downloaded from the Internet, where the resolutions of each image are considerably different. Part B is composed of 400 training images and 316 test images taken from streets in Shanghai, and the resolution of each image is 768×1024.

NWPU-Crowd [8], a large-scale and challenging dataset, consists of 5,109 images, 2,133,375 annotated heads with points and boxes in total, where these images are split into a training set (3,109), validation set (500), testing (1,500), respectively. The dataset has obvious strengths: negative samples, online testing for fair evaluation, higher resolution, most annotated head, and large appearance variation, compared with previous datasets in the real world.

UCF-CC-50 [30] only contains 50 images of intensely congested scenes, a very challenging crowd counting dataset,

TABLE II
THE CROP SIZE OF IMAGE PATCHES.

Datasets	(m, n)
ShanghaiPart A	(256, 256)
ShanghaiPart B	(512, 512)
UCF-CC-50	(256, 256)
UCF-QNRF	(512, 512)
NWPU-Crowd	(384, 384)

during a total of 63,974 head annotations. The number of pedestrians in each image ranges from 94 to 4,543, with an average number of 1,280 persons. We perform 5-fold cross-validation by randomly selecting images to train and test our proposed approach due to its small size.

UCF-QNRF [31] contains 1,535 images from the Internet with 1,251,642 annotations, where it is divided into train and test sets of 1,201 and 334 images, respectively. The number of pedestrians in each image varies from 49 to 12,865, with an average count of 815.4. Specifically, the image resolutions and scale varied dramatically comparing other datasets.

B. Loss Function

We utilize SmoothL1 loss to measure the difference between predicted count and ground truth count. Compared with L1, Smooth L1 is differentiable at the zero point and more robust to outliers. It is defined as:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \left\{ \begin{array}{l} 0.5(C_i^{ES} - C_i^{GT})^2, |C_i^{ES} - C_i^{GT}| < 1 \\ |C_i^{ES} - C_i^{GT}| - 0.5, \text{otherwise} \end{array} \right. \quad (10)$$

where N denotes the total number of images, and C_i^{ES} , C_i^{GT} represent the estimated count and ground truth count, respectively.

C. Training details

All experimental training and evaluation are implemented on the platform of PyTorch [32] with a GeForce RTX 3090 GPU. The CNN feature extraction module (CFM) of the proposed JCTNet is leveraged from the first ten layers of VGG16-BN [17]. The modified Swin Transformer blocks inspired by the [14] are used as the Transformer feature extraction module (TFM). The hyper-parameters of the TFM are set as follows:

- TFM: embed_dim = 256, window_size = 4, depths = [8, 8, 8], num_heads = [8, 8, 8, 8],

where embed_dim is the embedding dimension of hidden layers, depths is the depth of each modified Swin Transformer block, and num_heads is the number of attention heads in different layers. In the training phase, we randomly crop image patches with the size of $m \times n$ pixels from the original image to ensure our network can be multi-batch trained and promote performance at a lower time cost. The m, n configuration is detailed in Table 2. We use AdamW [33] optimizer with a learning rate of 1e-5 and weight decay of 1e-4 to train our model by minimizing the loss function Eq. (10). The batch size is set to 8 or 16 according to GPU computing power and the number of iterations is set to 2000.

V. EXPERIMENTS AND RESULTS

In this section, we display the evaluation metrics and compare the results of the proposed JCTNet with state-of-the-art methods. In the end, we perform extensive ablation experiments to validate the effectiveness of our method.

A. Evaluation metrics

To evaluate the accuracy of our approach, the mean absolute error (MAE), the mean squared error (MSE), and the mean normalized absolute error (NAE) [8] are adopted as metrics. Specifically, equations are defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^N |C_i^{ES} - C_i^{GT}| \quad (11)$$

$$MSE = (\frac{1}{N} \sum_{i=1}^N |C_i^{ES} - C_i^{GT}|^2)^{\frac{1}{2}} \quad (12)$$

$$NAE = \frac{1}{N} \sum_{i=1}^N \frac{|C_i^{ES} - C_i^{GT}|}{C_i^{GT}} \quad (13)$$

where N is the number of test images. C_i^{ES} is the estimated count of the i -th image, and C_i^{GT} is the corresponding ground truth count of the i -th image.

B. Comparisons with state-of-the-art

The elaborate comparisons and visual presentation are conducted with state-of-the-art on five benchmark datasets to demonstrate the effectiveness of the proposed weakly-supervised crowd counting approach JCTNet. We compare the proposed method with the existing fully-supervised methods (via location-level annotations) and weakly-supervised methods (via count-level annotations).

Result on ShanghaiTech. To evaluate the effectiveness of the proposed JCTNet, we compare the model with existing state-of-the-art 14 fully-supervised and 4 weakly-supervised methods on the ShanghaiTech dataset. As shown in Table 3, compared with the weakly-supervised crowd counting methods, the proposed JCTNet can achieve the lowest MAE of 62.8 and a comparable MSE of 95.6 on Part A and obtain the second places MAE of 7.2 and MSE of 11.5 on Part B. Compared with the fully-supervised crowd counting methods, our model still obtains relatively outstanding performance, though it is unfair to measure the fully-supervised methods and weakly-supervised methods.

Result on UCF-QNRF. Compared with other weakly supervised works, our model obtains the best MAE result with 90 and the second best MSE with 161. Besides, The model parameters of TransCrowd and CCTrans [16] are 86M and 104M, respectively, in Table 1, while the proposed JCTNet parameters are only 28M, achieving a reduction of 67% and 73%. Compared with the fully supervised works, our model still obtains significant counting results.

Result on UCF-CC-50. This is an extremely ingested crowd but small dataset, so we perform a 5-fold cross-validation for fairly evaluating the capability of the JCTNet. As shown in Table 5, the proposed method achieves the best

TABLE III
ESTIMATION ERRORS ON THE SHANGHAI TECH DATASET. * REPRESENTS THE WEAKLY-SUPERVISED METHOD.

Method	Publication	Label		Part A		Part B	
		Location	Count	MAE	MSE	MAE	MSE
SANet [34]	ECCV18	✓	✓	67.0	104.5	8.4	13.6
SFCN+ [7]	CVPR19	✓	✓	64.8	107.5	7.6	13.0
TEDNet [35]	CVPR19	✓	✓	64.2	109.1	8.2	12.8
DADNet [3]	MM19	✓	✓	64.2	99.8	8.8	13.5
Density CNN [36]	TMM19	✓	✓	63.1	106.3	9.1	16.3
HACNN [11]	TIP19	✓	✓	62.9	94.9	8.1	13.4
CAN [4]	CVPR19	✓	✓	62.3	100.0	7.8	12.2
PaDNet [37]	TIP19	✓	✓	59.2	98.1	8.1	12.2
PGCNet [38]	ECCV19	✓	✓	57.0	86.0	8.8	13.7
DM-Count [39]	NeurIPS20	✓	✓	59.7	95.7	7.4	11.8
ASNet [21]	CVPR20	✓	✓	57.7	90.1	—	—
P2PNet [40]	ICCV21	✓	✓	52.7	85.1	6.3	9.9
BCCT [12]	Arxiv21	✓	✓	53.1	82.2	7.3	11.3
CCTrans [16]	Arxiv21	✓	✓	52.3	84.9	6.2	9.9
Yang et al. [18]*	ECCV20	✗	✓	104.6	145.2	12.3	21.2
MATT [10]*	PR21	✗	✓	80.1	129.4	11.7	17.5
TransCrowd [9]*	Arxiv21	✗	✓	66.1	105.1	9.3	16.1
CCTrans [16]*	Arxiv21	✗	✓	64.4	95.4	7.0	11.5
JCTNet(ours) *	—	✗	✓	62.8	95.6	7.2	11.5

TABLE IV
ESTIMATION ERRORS ON THE UCF-QNRF DATASET. * REPRESENTS THE WEAKLY-SUPERVISED METHOD.

Method	Publication	Label		UCF-QNRF	
		Location	Count	MAE	MSE
SFCN+ [7]	CVPR19	✓	✓	102	171
TEDNet [35]	CVPR19	✓	✓	113	188
DADNet [3]	MM19	✓	✓	113	189
Density CNN [36]	TMM19	✓	✓	101	186
CAN [4]	CVPR19	✓	✓	107	183
PaDNet [37]	TIP19	✓	✓	96	170
DM-Count [39]	NeurIPS20	✓	✓	85	148
ASNet [21]	CVPR20	✓	✓	91	159
P2PNet [40]	ICCV21	✓	✓	85	154
BCCT [12]	Arxiv21	✓	✓	83	143
CCTrans [16]	Arxiv21	✓	✓	82	142
TransCrowd [9]*	Arxiv21	✗	✓	97	168
CCTrans [16]*	Arxiv21	✗	✓	92	158
JCTNet(ours) *	—	✗	✓	90	161

result with an MSE of 222.9 and MSE of 306.5, compared with other weakly supervised methods.

Result on NWPU-Crowd. Table 6 presents the counting errors on the NWPU-Crowd dataset. The proposed JCTNet wins second place with an MAE of 83.4 and MSE of 385.0 compared with the weakly supervised methods. The results of JCTNet are relatively poor compared with CCTrans [16], but it has about a quarter of the parameters of method CCTrans, only 28M as shown in Table 1. We find that the proposed model performs poorly on large-scale images containing more pedestrian annotations, suggesting that models constrained by count-level annotations cannot better adapt to scale changes. This will be the next research.

TABLE V
ESTIMATION ERRORS ON THE UCF-CC-50 DATASET. * REPRESENTS THE WEAKLY-SUPERVISED METHODS.

Method	Publication	Label		UCF-CC	
		Location	Count	MAE	MSE
SANet [34]	ECCV18	✓	✓	377.6	509.1
SFCN+ [7]	CVPR19	✓	✓	214.2	318.2
TEDNet [35]	CVPR19	✓	✓	249.4	354.5
Density CNN [36]	TMM19	✓	✓	244.6	341.7
CAN [4]	CVPR19	✓	✓	212.2	243.7
DM-Count [39]	NeurIPS20	✓	✓	211.0	291.5
ASNet [21]	CVPR20	✓	✓	174.8	251.6
P2PNet [40]	ICCV21	✓	✓	172.7	256.2
CCTrans [16]	Arxiv21	✓	✓	168.7	234.5
MATT [10]*	PR21	✗	✓	355.0	550.2
CCTrans [16]*	Arxiv21	✗	✓	245.0	343.6
JCTNet(ours) *	—	✗	✓	222.9	306.5

C. Visualization presentation

Visualization of convergence curves. Fig. 4 displays the convergence curves of the proposed JCTNet on the ShanghaiTech Part A dataset, including Smooth L1 loss, MAE, and MSE metrics, where the starting epoch of the three curves are from 0 to 2,000, 100 to 2,000, and 100 to 2,000, respectively. There are two observations: (1) The training loss can drop rapidly and converge fast. (2) The model has achieved the best counting performance at the 353-th epoch, i.e., MAE of 62.8 and MSE of 95.6. This figure further demonstrates the effectiveness of the proposed model.

Visualization of the model's components. For powerfully exploring what the model attends to in weakly supervision crowd counting, we present qualitative visualizations of intermediate feature maps of the model, consisting of the output feature of CFM, TFM, and CRM. Specially, for the features of size $H \times W \times C$ (height, width, channel), we select one

TABLE VI
ESTIMATION ERRORS ON THE NWPU-CROWD DATASET. * REPRESENTS THE WEAKLY-SUPERVISED METHODS.

Method	Publication	Label		Val		Test		
		Location	Count	MAE	MSE	MAE	MSE	NAE
MCNN [1]	CVPR16	✓	✓	218.5	700.6	232.5	714.6	1.063
SANet [34]	ECCV18	✓	✓	171.1	471.5	190.6	491.4	0.991
CSRNet [41]	CVPR18	✓	✓	104.9	433.4	121.3	387.8	0.604
SFCN+ [7]	CVPR19	✓	✓	95.5	608.3	105.7	424.1	0.254
CAN [4]	CVPR19	✓	✓	93.6	489.9	106.3	386.5	0.295
BL [42]	ICCV19	✓	✓	93.6	470.4	105.4	454.2	0.203
DM-Count [39]	NeurIPS20	✓	✓	70.5	357.6	88.4	388.6	0.169
KDMG [43]	TPAMI20	✓	✓	—	—	100.5	415.5	—
NoisyCC [44]	NeurIPS20	✓	✓	—	—	96.9	534.2	—
P2PNet [40]	ICCV21	✓	✓	—	—	77.4	362.0	—
BCCCT [12]	Arxiv21	✓	✓	53.0	170.3	82.0	366.9	0.164
CCTrans [16]	Arxiv21	✓	✓	38.6	87.8	69.3	299.4	0.135
TransCrowd [9]*	Arxiv21	✗	✓	88.4	400.5	117.7	451.0	0.244
CCTrans [16]*	Arxiv21	✗	✓	48.6	121.1	79.8	344.4	0.157
JCTNet(ours) *	—	✗	✓	68.1	310.6	83.4	385.0	0.185

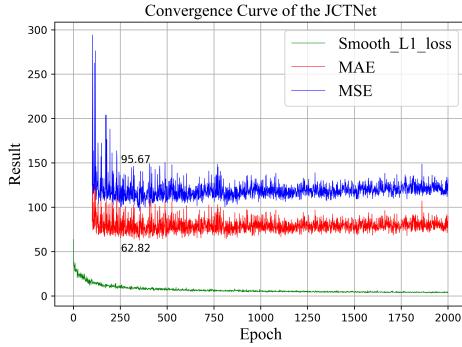


Fig. 4. The convergence curves of Smooth L1 Loss, MAE, and MSE metrics of the proposed JCTNet on the ShanghaiTech Part A dataset.

from the C feature maps for visualization, as shown in Fig. 5. We observe that the TFM can successfully learn the contrastive features and distinguish foreground and background regions (in black rectangles), and to our surprise, CRM can significantly focus on crowd areas (in red rectangles) and get superior counting results only with count-level annotations.

Visualization of the heat map of the last convolution layer of the proposed JCTNet. In Fig. 6, we give the high-quality heat map of the last convolution layer of the proposed JCTNet to further understand what the model prefers to focus on in the weakly supervision crowd counting, inspired by [45]. We observe that although the model is only weakly supervised by count-level annotations, it is still able to pay close attention to crowd regions.

For the above phenomenon, we give a possibility analysis: (1) Transformer component can sufficiently learn the contrastive features in favor of distinguishing various objects. (2) Then the CNN can count various objects with similar characteristics, respectively, due to its significant inductive bias capability. (3) Finally, the model regards the category whose count result is closest to the ground truth as the feature to take into count, i.e., the target feature. We hope this exploration will spur further research in the field.

D. Ablation Experiments

In this section, we conduct multigroup ablation experiments to evaluate the effectiveness of different components on the ShanghaiTech Part A dataset.

Ablation study on model's components. Table 7 presents the ablation study of the component of the proposed JCTNet on the ShanghaiTech Part A dataset. From the table, CFM, TFM, and CRM provide progressive improvements, respectively. Specifically, CFM achieves the MAE of 77.6 and the MSE of 127.5, and by adding TFM over CFM, we observe performance improvement of 13.4 and 21.8 on MAE and MSE, which demonstrates the effectiveness of the joint learning between CNN and Transformer. By further adding CRM, the best results are obtained.

TABLE VII
ABLATION STUDY ON KEY COMPONENTS OF JCTNET.

Method	Label		Part A	
	Location	Count	MAE	MSE
CFM	✗	✓	77.6	127.5
CFM + CRM	✗	✓	75.0	126.3
CFM + TFM	✗	✓	64.2	105.7
CFM + TFM + CRM	✗	✓	62.8	95.6

Ablation study on TFM's embedding dimension and Swin Transformer Layer (STL) number of each Modified Swin Transformer Block (MSTB). Table 8 the ablation study of embedding dimension and STL number of each MSTB for the proposed JCTNet on ShanghaiTech A. The x, y of the TFM (x, y) denote the embedding dimension and STL number of each MSTB, respectively. From the table, by enlarging the embedding dimension and the number of STLs, the count error (MAE) decreases gradually as more parameters are involved.

Ablation study on the Loss function. We perform quantitative analysis between the L1 loss and Smooth L1 loss on the ShanghaiTech Part A, as shown in Table 9. The MAE and MSE can be reduced from 63.7 and 98.1 to 62.8 and 95.6 by employing the Smooth L1 loss.

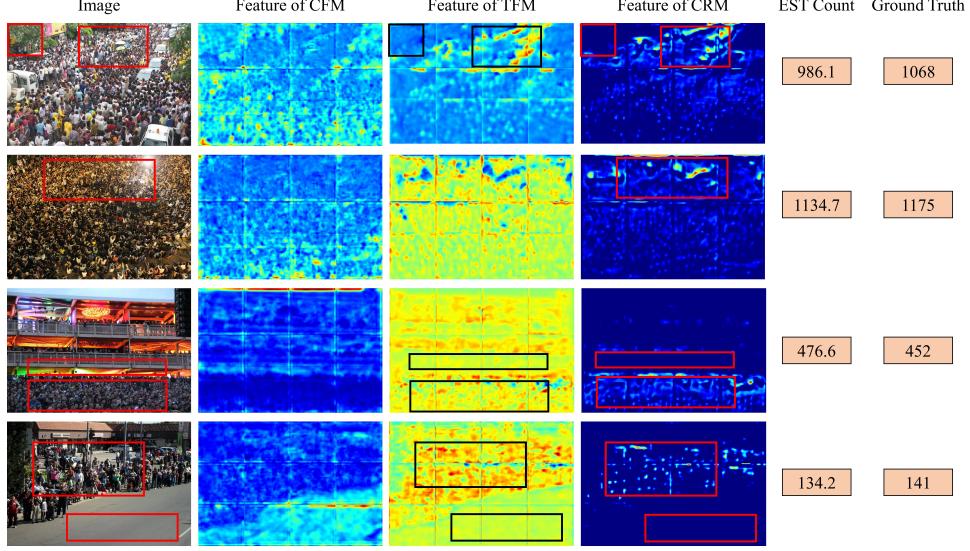


Fig. 5. The visualization of the feature maps: the output feature of CFM, TFM, and CRM. For weakly supervised crowd counting, the TFM of the proposed JCTNet can be still able to distinguish crowd region and background (the black rectangle), and then the CRM focus on counting crowd region (the red rectangle).

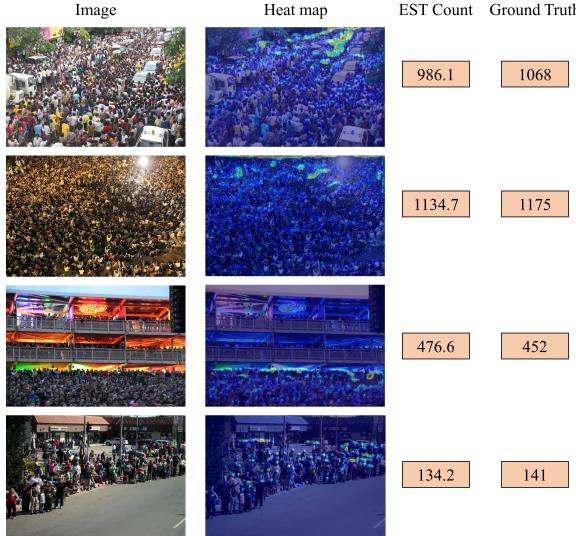


Fig. 6. The visualization of the heat map of the last convolution layer of the proposed JCTNet on input images: It can be observed that the model is still able to focus the crowd regions for weakly supervised crowding counting.

VI. CONCLUSION

In this paper, we proposed a joint CNN and Transformer learning network for the efficient weakly supervised crowd counting, named JCTNet. The model includes three parts: CNN feature extraction module (CFM) for extracting crowd semantic information, Transformer extraction module (TFM) for capturing the global context and thoroughly learning the contrast features between foreground and background, and counting regression module (CRM) for estimating the final count of people. Extensive experiments and visualization demonstrate the effectiveness of the proposed JCTNet on five mainstream datasets.

TABLE VIII
ABLATION STUDY ON THE EMBEDDING DIMENSION AND NUMBER OF STL OF TFM.

Method	Label		Part A		Parameters (M)
	Location	Count	MAE	MSE	
CFM + TFM (64, 8) + CRM	✗	✓	66.3	111.2	10.4
CFM + TFM (128, 8) + CRM	✗	✓	64.5	99.9	14.2
CFM + TFM (256, 8) + CRM	✗	✓	62.8	95.6	28.8
CFM + TFM (256, 2) + CRM	✗	✓	68.4	117.1	16.3
CFM + TFM (256, 4) + CRM	✗	✓	65.7	102.8	20.5
CFM + TFM (256, 8) + CRM	✗	✓	62.8	95.6	28.8

TABLE IX
ABLATION STUDY ON THE LOSS FUNCTION.

Method	Label		Part A	
	Location	Count	MAE	MSE
L1 Loss	✗	✓	63.7	98.1
Smooth L1 Loss	✗	✓	62.8	95.6

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No. 61971073).

REFERENCES

- [1] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 589–597.
- [2] L. Boominathan, S. S. Kruthiventi, and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 640–644.
- [3] D. Guo, K. Li, Z.-J. Zha, and M. Wang, "Dadnet: Dilated-attention-deformable convnet for crowd counting," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1823–1832.
- [4] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5094–5103.

- [5] L. Zhu, Z. Zhao, C. Lu, Y. Lin, and Y. Peng, "Dual path multi-scale fusion networks with attention for crowd counting," *arXiv preprint arXiv:1902.01115*, 2019.
- [6] G. Schröder, T. Senst, and E. Bochinski, "Optical flow dataset and benchmark for visual crowd analysis," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2018, pp. 1–6.
- [7] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8190–8199.
- [8] Q. Wang, J. Gao, W. Lin, and X. Li, "Nwpu-crowd: A large-scale benchmark for crowd counting and localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [9] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai, "Transcrowd: Weakly-supervised crowd counting with transformer," *arXiv preprint arXiv:2104.09116*, 2021.
- [10] Y. Lei, Y. Liu, and P. Zhang, "Towards using count-level weak supervision for crowd counting," *Pattern Recognition*, vol. 109, p. 107616, 2021.
- [11] V. A. Sindagi and V. M. Patel, "Ha-ccn: Hierarchical attention-based crowd counting network," *IEEE Transactions on Image Processing*, vol. 29, pp. 323–335, 2020.
- [12] G. Sun, Y. Liu, T. Probst, and D. Paudel, "Boosting crowd counting with transformers," *arXiv preprint arXiv:2105.10926*, 2021.
- [13] A. Dosovitskiy, L. Beyer, and A. Kolesnikov, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [14] Z. Liu, Y. Lin, Y. Cao, and H. Hu, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.
- [15] X. Chu, Z. Tian, Y. Wang, and B. Zhang, "Twins: Revisiting the design of spatial attention in vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [16] Y. Tian, X. Chu, and H. Wang, "Cctrans: Simplifying and improving crowd counting with transformer," *arXiv preprint arXiv:2109.14483*, 2021.
- [17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [18] Y. Yang, G. Li, Z. Wu, and L. Su, "Weakly-supervised crowd counting learns from sorting rather than locations," in *European Conference on Computer Vision*. Springer, 2020, pp. 1–17.
- [19] G. Gao, J. Gao, and Q. Liu, "Cnn-based density estimation and crowd counting: A survey," *arXiv preprint arXiv:2003.12783*, 2020.
- [20] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4031–4039.
- [21] X. J. et al, "Attention scaling for crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4706–4715.
- [22] D. Sam, N. Sajjan, and H. Maurya, "Almost unsupervised learning for dense crowd counting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8868–8875.
- [23] Y. Li, K. Zhang, J. Cao, and R. Timofte, "Localvit: Bringing locality to vision transformers," *arXiv preprint arXiv:2104.05707*, 2021.
- [24] H. Cao, Y. Wang, J. Chen, and D. Jiang, "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537*, 2021.
- [25] N. Carion, F. Massa, G. Synnaeve, and N. Usunier, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.
- [26] R. Faster, "Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 9199, no. 10.5555, pp. 2969–2969 250, 2015.
- [27] X. Ding, X. Zhang, N. Ma, and J. Han, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13733–13742.
- [28] V. K. Valloli and K. Mehta, "W-net: Reinforced u-net for density map estimation," *arXiv preprint arXiv:1903.11249*, 2019.
- [29] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.
- [30] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2547–2554.
- [31] H. I. et al, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 532–546.
- [32] A. P. et al, "Automatic differentiation in pytorch," 2017.
- [33] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
- [34] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.
- [35] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, and L. Shao, "Crowd counting and density estimation by trellis encoder-decoder networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6126–6135.
- [36] X. Jiang, L. Zhang, T. Zhang, P. Lv, B. Zhou, Y. Pang, M. Xu, and C. Xu, "Density-aware multi-task learning for crowd counting," *IEEE Transactions on Multimedia*, vol. 23, pp. 443–453, 2021.
- [37] Y. Tian, Y. Lei, J. Zhang, and J. Z. Wang, "Padnet: Pan-density crowd counting," *IEEE Transactions on Image Processing*, vol. 29, pp. 2714–2727, 2020.
- [38] Z. Yan, Y. Yuan, W. Zuo, X. Tan, Y. Wang, S. Wen, and E. Ding, "Perspective-guided convolution networks for crowd counting," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 952–961.
- [39] B. Wang, H. Liu, and D. Samaras, "Distribution matching for crowd counting," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1595–1607, 2020.
- [40] Q. Song, C. Wang, and Z. Jiang, "Rethinking counting and localization in crowds: A purely point-based framework," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3365–3374.
- [41] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1091–1100.
- [42] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6142–6151.
- [43] J. Wan, Q. Wang, and A. Chan, "Kernel-based density map generation for dense object counting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [44] J. Wan and A. Chan, "Modeling noisy annotations for crowd counting," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3386–3396, 2020.
- [45] R. Selvaraju, M. Cogswell, and A. Dasv, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.