Spatio-channel Attention Blocks
for Cross-modal Crowd Counting

Youjia Zhang[0000−0002−5671−3232], Soyun Choi[0000−0002−2701−4782], and
Sungeun Hong[0000−0003−1774−9168]

Department of Electrical and Computer Engineering, Inha University, South Korea
{zhangyoujia,sychoi}@inha.edu, csehong@inha.ac.kr

Abstract. Crowd counting research has made significant advancements
in real-world applications, but it remains a formidable challenge in cross-
modal settings. Most existing methods rely solely on the optical features
of RGB images, ignoring the feasibility of other modalities such as ther-
mal and depth images. The inherently significant differences between the
different modalities and the diversity of design choices for model archi-
tectures make cross-modal crowd counting more challenging. In this pa-
per, we propose Cross-modal Spatio-Channel Attention (CSCA) blocks,
which can be easily integrated into any modality-specific architecture.
The CSCA blocks first spatially capture global functional correlations
among multi-modality with less overhead through spatial-wise cross-
modal attention. Cross-modal features with spatial attention are sub-
sequently refined through adaptive channel-wise feature aggregation. In
our experiments, the proposed block consistently shows significant per-
formance improvement across various backbone networks, resulting in
state-of-the-art results in RGB-T and RGB-D crowd counting.

1 Introduction

Crowd counting is a core and challenging task in computer vision, which aims
to automatically estimate the number of pedestrians in unconstrained scenes
without any prior knowledge. Driven by real-world applications, including traf-
fic monitoring [1], social distancing monitoring [16], and other security-related
scenarios [56], crowd counting plays an indispensable role in various fields. Over
the last few decades, an increasing number of researchers have focused on this
topic, achieving significant progress [51,63,60,50] in counting tasks.

In general, the various approaches for crowd counting can be roughly di-
vided into two categories: detection-based methods and regression-based meth-
ods. Early works [8,22,10,46,12,54] on crowd counting mainly use detection-based
approaches, which detect human heads in an image as can be seen in drone ob-
ject detection tasks [9,18]. In fact, most detection-based methods assume that
the crowd in the image is made up of individuals that can be detected by the
given detectors. There are still some significant limitations for tiny heads in
higher density scenarios with overlapping crowds and serious occlusions. To re-
duce the above problems, some researchers [6,7,19,40] introduce regression-based

(a) RGB-T (RGB<T)   (b) RGB-T (RGB>T)   (c) RGB-D (RGB<D)   (d) RGB-D (RGB>D)
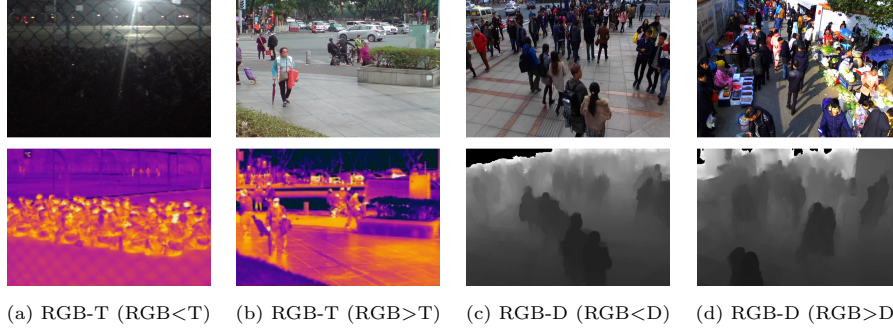
Fig. 1. Visualization of RGB-T and RGB-D pairs. (a) shows the positive effects of thermal images in extremely poor illumination. (b) shows the negative effects caused by additional heating objects. In case (c), the depth image provides additional information about the position and size of the head, but in scene (d), part of the head information in the depth image is corrupted by cluttered background noise.

methods where they view crowd counting as a mapping problem from an image to the count or the crowd-density map. In this way, the detection problem can be circumvented. Recent works [64,26,59,44] have shown the significant success of density map estimation in RGB crowd counting, and it has become the mainstream method for the counting task in complex crowd scenes.

Despite substantial progress, crowd counting remains a long-standing challenging task for various situations such as occlusion, high clutter, non-uniform lighting, non-uniform distribution, etc. Due to the limitations of methods or insufficient data sources, most previous works have focused on the RGB crowd counting and failed to obtain high-quality density maps in unconstrained scenarios. RGB images can provide strong evidence in bright illumination conditions while they are almost invisible in the dark environment. Recently, some researchers [31,33] indicate that the thermal feature helps better distinguish various targets in poor illumination conditions. Therefore, as shown in Fig. 1-(a), thermal images can be used to improve the performance of crowd counting tasks. However, from Fig. 1-(b), we can find that additional heating objects seriously interfere with the thermal image.

Furthermore, besides exploiting various contextual information from the optical cues, the depth feature [3,55,27] has recently been utilized as supplementary information of the RGB feature to generate improved crowd density maps. From Fig. 1-(c), we observe that crowd movement leads to aggregation behavior and large variation in head size. Thus, it is insufficient to generate pixel-wise density maps using only RGB data. Depth data naturally complements RGB information by providing the 3D geometry to 2D visual information, which is robust to varying illumination and provides additional information about the localization and size of heads. However, as shown in Fig. 1-(d),the depth information would be critically affected by noise under the condition of cluttered background.

Despite the many advantages of using additional modalities besides RGB in crowd counting, there are few related studies in crowd counting. In this study, we

focus on two main issues that need to be addressed for a more robust cross-modal crowd counting as follows:

i Crowd counting data is obtained from the different modalities, which means that they vary significantly in distribution and properties. Since there are substantial variations between multimodal data, how to effectively identify their differences and unify multiple types of information is still an open problem in cross-modal crowd counting. Additionally, since additional modality other than RGB is not fixed (either thermal or depth), it is very critical to adaptively capture intra-modal and inter-modal correlation.

ii For cross-modal crowd counting, various types can be given as additional modalities besides RGB. At this time, constructing a cross-modal model architecture from scratch usually entails a lot of trial and error. We have seen the well-explored model architectures show promising results in various fields through transfer learning or model tuning. In this respect, if the widely used RGB-based unimodal architecture can be extended to multimodal architecture without bells and whistles, we can relieve the effort of design choice of the model architecture.

In this paper, we propose a novel plug-and-play module called Cross-modal Spatio-Channel Attention (CSCA) block consisting of two main modules. First, Spatial-wise Cross-modal Attention (SCA) module utilizes an attention mechanism based on the triplet of 'Query', 'Key', and 'Value' widely used in non-local-based models [53,66,58]. Notably, conventional non-local blocks are unsuitable for cross-modal tasks due to their high computational complexity and the fact that they are designed to exploit self-attention within a single modality. To address these issues, SCA module is deliberately designed to understand the correlation between cross-modal features, and reduces computational overhead through the proposed feature re-assembling scheme. Given an input feature map with size C×H×W, the computational complexity of a typical non-local block is $O(CH^2W^2)$ while our SCA block can reduce the complexity by a factor of G_l to $O(CH^2W^2/G_l)$. Here, the G_l refers to the spatial re-assembling factor in the lth block of the backbone network.

Importantly, we note that non-local blocks widely used in various vision tasks focus on spatial correlation of features but do not directly perform feature recalibration at the channel level. To supplement the characteristics of non-local-based blocks considering only spatial correlation, our Channel-wise Feature Aggregation (CFA) module adaptively adjusts the spatially-correlated features at the channel level. Unlike previous works, which use pyramid pooling feature differences to measure multimodal feature variation [31] or fine-tune the density map [23], our SCA aims to spatially capture global feature correlations between multimodal features to alleviate misalignment in multimodal data. Moreover, instead of adding auxiliary tasks for other modal information [3,55,27], our method dynamically aggregate complementary features along the channel dimension by the CFA block. We show that the proposed two main modules complement each other on two challenging RGB-T [31] and RGB-D [27]. Additionally, we conduct

various ablation studies and comparative experiments in cross-modal settings that have not been well explored. In summary, the major contributions of this work are as follows:

- We propose a new plug-and-play module that can significantly improve performance when integrated with any RGB-based crowd counting model.

- Our cross-modal attention with feature re-assembling is computationally efficient and has higher performance than conventional non-local attention. Moreover, our channel-level recalibration further improves the performance.

- The proposed CSCA block consistently demonstrates significant performance enhancements across various backbone networks and shows state-of-the-art performance in RGB-T and RGB-D crowd counting.

- We extensively conduct and analyze comparative experiments between multi-modality and single-modality, which have not been well explored in existing crowd counting, and this can provide baselines for subsequent studies.

2   Related Works

2.1   Detection-based methods

Early works of crowd counting focus on detection-based approaches. The initial method [28,49] mainly employs detectors based on hand-crafted features. The typical traditional methods [8,22,10,24] train classifiers to detect pedestrians using wavelet, HOG, edge, and other features extracted from the whole body, which degrades seriously for those very crowded scenes. With the increase of crowd density, the occlusion between people becomes more and more serious. Therefore, methods [12,54] based on partial body detection (such as head, shoulders, etc.) are proposed to deal with the crowd counting problem. Compared with the overall detection, this method has a slight improvement in effect. Recently, detection-based [27,30,47,2] methods have also been further improved with the introduction of more advanced object detectors. For example, some works [30,27] design specific architectures such as Faster RCNN [39] or RetinaNet [29] to promote the performance of crowd counting by improving the robustness of detectors for tiny head detection. But, they still present unsatisfactory results on challenging images with very high densities, significant occlusions, and background clutter in extremely dense crowds, which are quite common in crowd counting tasks.

2.2   Regression–based methods

Some of the recent approaches [7,64,26,15,52,4,43,32] attempt to tackle this problem by bypassing completely the detection problem. They directly address it as a mapping problem from images to counts or crowd density maps. Recently, numerous models utilize basic CNN layers [15,52], multi-column based models [64,4],

or single-column based models [26,34,5] to extract rich information for density estimation and crowd counting. To capture multi-scale information at different resolutions, MCNN [64] proposes a three branches architecture. And, CrowdNet [4] combines low-level features and high-level semantic information. To reduce the information redundancy caused by the multi-column method, SANet [5] is built on the Inception architecture. And, CSRNet [26] adopts dilated convolution layers to expand the receptive field. Regression-based methods capture low-level features and high-level semantic information through multi-scale operations or dilated convolution, which show remarkable success in crowd counting tasks. However, most methods only focus on the RGB images and may fail to accurately recognize the semantic objects in unconstrained scenarios.

2.3   Multi-modal learning for crowd counting

Recently, multimodal feature fusion has been widely studied across various fields [17,45,36]. The cross-modal feature fusion aims to learn better feature representations by exploiting the complementarity between multiple modalities. Some cross-modal fusion methods [14,45] fuse the RGB source and the depth source in the "Early Fusion" or "Late Fusion" way. Besides, inspired by the success of multi-task learning in various computer vision tasks, some researchers [65,42] utilize the depth feature as the auxiliary information to boost the performance on the RGB crowd counting task. Similarly, the perspective information [41] during the density estimation procedure is also treated as auxiliary data to improve the performance of RGB crowd counting. However, these methods rely on handcrafted features instead of fusing multimodal data with predicting weights from different modalities. Moreover, some researchers [27,13] combine the depth information into the detection procedure for the crowd counting task, which would be limited by the performance of the detectors. The IADM [31] proposes to collaboratively represent the RGB and thermal images. However, it does not fully consider the different characteristics and feature space misalignment of RGB data and thermal imaging data. Different from previous work, in this work, we propose a novel cross-modal spatio-channel attention block, which adaptively aggregates cross-modal features based on both spatial-wise and channel-wise attention. Furthermore, our SCA block is robust to cross-modal feature variation at lower complexity compared to the conventional non-local block.

3   Proposed Method

3.1   Overview

We proposed a unified framework to extend the existing baseline models from unimodal crowd counting to the multimodal scene. As shown in Fig. 2, the framework for cross-modal crowd counting consists of two parts: modality-specific branches and the Cross-modal Spatio-Channel Attention (CSCA) block. Given pairs of multimodal data like RGB-T or RGB-D pairs, we first utilize the backbone network (*e.g.*, BL [34], CSRNet [26], etc.) to extract the modality-specific
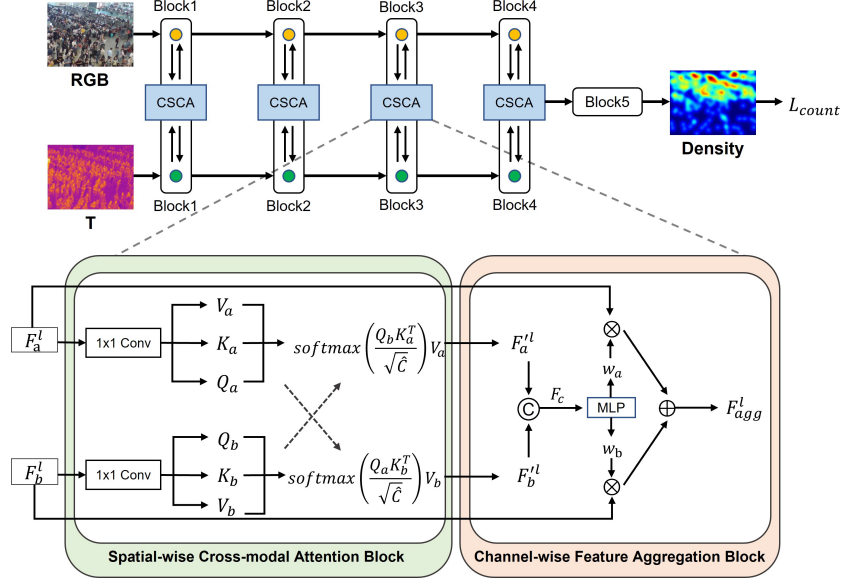
Fig. 2. The architecture of the proposed unified framework for extending existing baseline models from unimodal crowd counting to multimodal scenes. Our CSCA module is taken as the cross-modal solution to fully exploit the multimodal complementarities. Specifically, the CSCA consists of SCA to model global feature correlations among multimodal data, and CFA to dynamically aggregate complementary features.

features of RGB images and thermal images. To fully exploit the multimodal complementarities, we embed CSCA after different convolutional blocks of the two modality-specific branches to further update the feature maps and get the complementary fusion information.

To ensure informative feature interaction between modalities, our CSCA consists of two operations: Spatial-wise Cross-modal Attention (SCA) block and Channel-wise Feature Aggregation (CFA) block, as illustrated in the bottom part of Fig. 2. The SCA block utilizes an improved cross-modal attention block to model the global feature correlations between RGB images and thermal/depth images, which has a larger receptive field than the general CNN modules and facilitates complementary aggregation operation. And, the CFA block adaptively extracts and fuses cross-modal information based on channel correlations. Consequently, the framework can capture the non-local dependencies and aggregate the complementary information of cross-modal data to expand from unimodal crowd counting to multimodal crowd counting.

3.2   Spatial-wise Cross-modal Attention (SCA)

SCA module aims to capture global feature correlations among multimodal data with less overhead based on cross-modal attention. Recently, the Transformer-based architecture [48] has achieved comparable or even superior performance on computer vision tasks. The multi-head attention as one of the most important
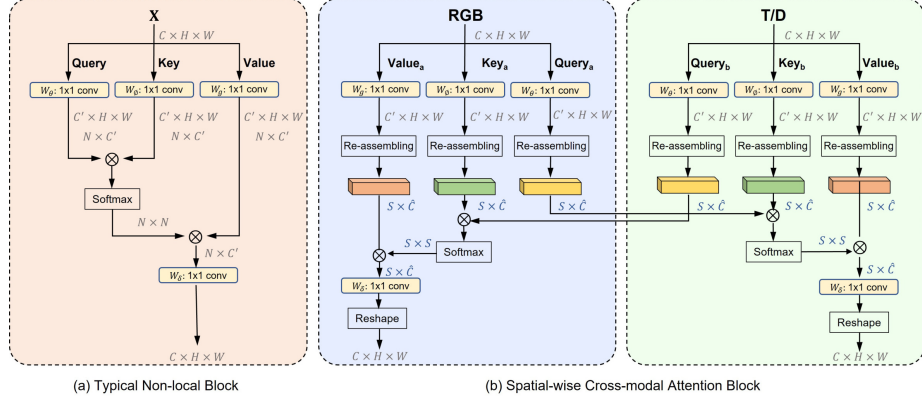
Fig. 3. Comparison between block structures (a) typical non-local block (b) proposed SCA block where the blue font indicates dimension transformation of SCA block, $\hat{C} = C' G_l$ and $S \times \hat{C} = N \times C'$ while $S = N/G_l$.

parts of the Transformer consists of several non-local attention layers running in parallel. A typical non-local block [53] is shown in Fig. 3-(a), which aims to capture the global relationships among all locations of the input image. Consider an input feature map $X \in \mathcal{R}^{C \times H \times W}$, where C, H, and W indicate the channel number, spatial height, and width, respectively. The feature map is first projected and flattened into three vectors: Query $Q \in \mathcal{R}^{N \times C'}$, Key $K \in \mathcal{R}^{N \times C'}$ and Value $V \in \mathcal{R}^{N \times C'}$ by three different 1×1 convolutions, where $N = H \times W$, and C' refers to the new embedding channel, $C' = C/2$. Then, the triplet of Query, Key, and Value are used to encode the relationship at any position in the feature map, thereby capturing the global context information. The output can be computed as a weighted sum of values where the weight is assigned to each value by the compatibility of the query with the corresponding key. This procedure can be computed efficiently with matrix multiplications as:

$$Z = f(QK^T)V, \tag{1}$$

where the $f(\cdot)$ is a normalization function, such as softmax, rescaling, etc. Finally, the non-local block recovers the channel dimension from C' to C by an additional 1×1 convolution layer.

However, the typical non-local block can only work on unimodal data, and the high complexity introduced by matrix multiplication hinders its use. To alleviate these problems, we propose the SCA block. Different from the typical non-local block, the input of our SCA Block is a pair of images with the different modalities. And it's Key and Value are from the same modality, while the Query is from another modality. For example, the Key and Value are from the RGB image modality, and the Query is generated from thermal image or depth image modality, and vice versa.

$$Z_a = softmax(\frac{Q_b K_a^T}{\sqrt{\hat{C}}})V_a, \tag{2}$$

where the a means the RGB modality and b means another modality (*i.e.*, thermal or depth modality). The output of the SCA block (Fig. 3-(b)), is calculated by multiplying Value with attention weights, which presents the similarity between the corresponding Query in modality b and all the Keys in another modality a. In other words, the SCA block aggregates and aligns the information from two modalities.

Thanks to the great power of cross-attention, many works [66,35,25,57] have utilized it to solve the misalignment. However, the matrix multiplication operation is very time and memory-consuming compared to convolutions and activation blocks. Some researchers [31] alleviate the problem by reducing the dimensions of the Key and Value through the pooling operation. But, it is hard to ensure that the performance would not degenerate too much by losing a lot of global information. Instead of the pooling function, we perform re-assembling operations in the SCA block to randomly divide the feature map into G_l patches in the spatial dimension, where l means the lth block of the backbone network. Thus, the Key, Value, and Query vectors in our SCA block become $K, V, Q \in \mathcal{R}^{S \times \hat{C}}$, where $S = HW/G_l$, $\hat{C} = C^{'} G_l$. Finally, we recover the channel dimension and reshape the result to $C \times H \times W$ size. Therefore, the computational cost of our SCA block is $O(CH^2W^2/G_l)$, which is G_l times less than the computational complexity $O(CH^2W^2)$ of a typical non-local block.

3.3   Channel-wise Feature Aggregation (CFA)

The SCA block can capture global features at all locations in multimodal images by spatial-wise attention. To make full use of the complementarity of multimodal data, we need to complementarily aggregate cross-modal features according to their characterization capabilities. To achieve this, we propose a CFA block to fusion the cross-modal features and capture channel-wise interactions. As shown in the bottom part of Fig. 2, we first concatenate the outputs of the SCA block to get the F_c. We then use an MLP layer and softmax function to learn the weight vectors $w_a, w_b \in \mathcal{R}^{C \times H \times W}$, which are used to re-weight the RGB feature and thermal/depth feature across the channels.

$$[w_a, w_b] = softmax(MLP(F_c)). \tag{3}$$

To fully exploit the aggregated information and suppress feature noise and redundancy simultaneously, we fusion the cross-modal features using the two channel-wise adaptive weights. The final aggregation feature map F_{agg}^l at l^{th} block can be obtained by the following formula:

$$F_{agg}^l = w_a * F_a^l + w_b * F_b^l. \tag{4}$$

Through the SCA block, the network can increase sensitivity to informative features on the important channels and enhance the final feature map. In practice, for each layer l, the updated feature maps $f_a^l = (F_{agg}^l + F_a^l)/2$ and $f_b^l = (F_{agg}^l + F_b^l)/2$ are propagated to the next blocks except the last layer.

Hence, the module can adaptively provide more discriminative and richer characteristics to increase cross-modal crowd counting performance.

4   Experiments

To evaluate our method, we carry out detailed experiments on two cross-modal crowd counting datasets: RGBT-CC benchmark [31] and ShanghaiTechRGBD [27]. And we also provide a thorough ablation study to validate the key components of our approach. Experimental results show that the proposed method considerably improves the performance of the cross-modal crowd counting.

4.1   Implementation details

In this work, our code is based on PyTorch [38]. Following [31], our model is implemented with three backbone networks: BL [34], MCNN [64], and CSRNet [26]. To make fair comparisons, we maintain a similar number of parameters to the original backbone models, and the same channel setting with the work [31], which takes 70%, 60%, and 60% of the original values.

For the input size, we feed the 640×480 RGB-T pairs and 1920×1080 RGB-D pairs, respectively. Notably, the original loss function of the adopted backbone network is used to train our framework. During training, our framework is optimized by Adam [21] with a 1e-5 learning rate. All the experiments are conducted using $1\times$ RTX A6000 GPU.

Datasets: The popular datasets, RGBT-CC [31] and ShanghaiTechRGBD [27], in the cross-modal crowd counting field are used in this study. The RGBT-CC dataset is a new challenging benchmark for RGB-T crowd counting, including 2,030 RGB-thermal pairs with resolution 640×480. All the images are divided into three sets: training set with 1030 pairs, validation set with 200 pairs, and testing set with 800 pairs, respectively. The ShanghaiTechRGBD dataset is a large-scale RGB-D dataset, which consists of 1193 training RGB-D image pairs and 1000 test RGB-D image pairs with a fixed resolution of 1920×1080.

Evaluation Metrics: As commonly used in previous works, we adopt the Root Mean Square Error (RMSE) and the Grid Average Mean absolute Error (GAME) as the evaluation metrics.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(P_i - \hat{P}_i\right)^2}$$

$$GAME(L) = \frac{1}{N} \sum_{i=1}^{N} \sum_{l=1}^{4^L} |P_i^l - \hat{P}_i^l|, \tag{5}$$

where N means numbers of testing images, P_i and $\hat{P}_i$ refer to the total count of the estimated density map and corresponding ground truth, respectively. Different from the RMSE, the GAME divides each image into 4^L non-overlapping

Table 1. The performance of our model implemented with different backbones on RGBT-CC.

Backbone	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
MCNN[64]	21.89	25.70	**30.22**	**37.19**	37.44
MCNN+CSCA	**17.37**	**24.21**	30.37	37.57	**26.84**
CSRNet[26]	20.40	23.58	**28.03**	**35.51**	35.26
CSRNet+CSCA	**17.02**	**23.30**	29.87	38.39	**31.09**
BL[34]	18.70	22.55	26.83	34.62	32.64
BL+CSCA	**14.32**	**18.91**	**23.81**	**32.47**	**26.01**

Table 2. The performance of our model implemented with different backbones on ShanghaiTechRGBD.

Backbone	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
BL[34]	8.94	11.57	15.68	22.49	12.49
BL+CSCA	**5.68**	**7.70**	**10.45**	**15.88**	**8.66**
CSRNet[26]	4.92	6.78	9.47	13.06	7.41
Ours	**4.39**	**6.47**	**8.82**	**11.76**	**6.39**

regions to measure the counting error in each equal area, where P_i^l and $\hat{P}_i^l$ are the estimated count and ground truth in a region l of image i. Note that $GAME(0)$ is equivalent to Mean Absolute Error (MAE).

4.2   Effectiveness of plug-and-play CSCA blocks

To evaluate the effectiveness of the proposed plug-and-play CSCA module, we easily incorporate CSCA as a plug-and-play module into backbone networks (*e.g.*, MCNN, CSRNet, and BL), and extend existing models from unimodal to multimodal on the RGBT-CC dataset. We report the result in Table 1. Note that for all baseline methods in the table, we estimate crowd counts by feeding the concatenation of RGB and thermal images. As can be observed, almost all instances of CSCA based method significantly outperform the corresponding backbone networks. Especially, our method far outperforms MCNN by 4.52 in MAE and 10.60 in RMSE. Compared to the CSRNet backbone, our CSCA provides a 3.88 and 4.17 improvement in MAE and RMSE, respectively. For the BL backbone, our method surpasses the baseline on all evaluation metrics.

Similarly, we also apply CSCA to the RGB-D crowd counting task. Furthermore, previous research [31] showed that MCNN is inefficient on the ShanghaiTechRGBD dataset with high-resolution images due to its time-consuming multi-column structure. So here we only compare the two backbone networks: BL and CSRNet. As shown in Table 2, our framework consistently outperforms their corresponding backbone networks on all evaluation metrics. In conclusion, all experimental results confirm that CCIM is universal and effective for cross-modal crowd counting.

Table 3. Counting performance evaluation on RGBT-CC.

Backbone	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
UCNet [61]	33.96	42.42	53.06	65.07	56.31
HDFNet [37]	22.36	27.79	33.68	42.48	33.93
BBSNet [11]	19.56	25.07	31.25	39.24	32.48
MVMS [62]	19.97	25.10	31.02	38.91	33.97
MCNN+IADM [31]	19.77	23.80	28.58	35.11	30.34
CSRNet+IADM [31]	17.94	21.44	26.17	33.33	30.91
BL+IADM [31]	15.61	19.95	24.69	32.89	28.18
CmCaF [23]	15.87	19.92	24.65	**28.01**	29.31
Ours	**14.32**	**18.91**	**23.81**	32.47	**26.01**

Table 4. Counting performance evaluation on ShanghaiTechRGBD.

Backbone	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
UCNet [61]	10.81	15.24	22.04	32.98	15.70
HDFNet [37]	8.32	13.93	17.97	22.62	13.01
BBSNet [11]	6.26	8.53	11.80	16.46	9.26
DetNet [30]	9.74	-	-	-	13.14
CL [20]	7.32	-	-	-	10.48
RDNet [44]	4.96	-	-	-	7.22
MCNN+IADM [31]	9.61	11.89	15.44	20.69	14.52
BL+IADM [31]	7.13	9.28	13.00	19.53	10.27
CSRNet+IADM [31]	**4.38**	**5.95**	**8.02**	**11.02**	7.06
Ours	4.39	6.47	8.82	11.76	**6.39**

4.3   Comparison with previous methods

In addition, to evaluate that our proposed model can be more effective and efficient, we compare our method with state-of-the-art methods on RGB-T and RGB-D crowd counting tasks. For fair comparisons, we directly use the optimal results of the comparison methods in the original papers. As shown in Table 3, compared with multimodal models, including UCNet [61], HDFNet [37], and BBSNet [62], our method outperforms these models by a wide margin on all evaluation metrics. Moreover, our model also achieves better performance on the RGBT-CC dataset compared to specially designed crowd counting models, including IADM [31], MVMS [11], and CmCaF [23]. Especially, our CSCA implemented with the BL backbone achieves state-of-the-art performance on almost all evaluation metrics.

For the RGB-D crowd counting task, we also compare two categories of methods including multimodal methods and specially-designed models for cross-modal crowd counting. For fair comparisons, we use them to estimate the crowd counts by following the work [31]. The performance of all comparison methods is shown in Table 4. The results of the proposed method on the ShanghaiTechRGBD dataset are encouraging, as we used the same parameters as our experiments on the RGBT-CC dataset and did not perform fine-tuning. Moreover, our model implemented with the CSRNet backbone outperforms most ad-

Table 5. Ablation study using the BL backbone on RGBT-CC.

Method	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
RGB	32.89	38.81	44.49	53.44	59.49
T	17.80	22.88	28.50	37.30	30.34
Early fusion	18.25	22.35	26.78	34.96	37.71
Late fusion	16.63	20.33	24.81	32.59	32.66
NL	17.91	21.12	25.06	32.61	32.60
SCA	17.85	21.13	25.07	32.56	30.60
CFA	17.78	21.36	25.69	33.14	32.37
NL+CFA	16.69	20.75	25.63	33.92	29.39
CSCA (SCA+CFA)	**14.32**	**18.91**	**23.81**	**32.47**	**26.01**

Table 6. Ablation study using the BL backbone on ShanghaiTechRGBD. Since methods based on non-local blocks run out of memory, we do not report those results.

Method	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
RGB	6.88	8.62	11.86	18.31	10.44
D	15.17	17.92	22.59	30.76	23.65
Early fusion	7.04	8.74	11.82	18.33	10.62
Late fusion	6.55	8.01	10.45	15.43	9.93
SCA	6.94	8.19	10.48	15.91	10.61
CFA	6.43	8.07	10.80	16.12	9.86
CSCA (SCA+CFA)	**5.68**	**7.70**	**10.45**	**15.88**	**8.66**

vanced methods and achieves a new state-of-the-art result of 6.39 in RMSE, the main evaluation metric for crowd counting.

4.4 Ablation studies

To verify the effectiveness of the proposed universal framework for cross-modal crowd counting, we investigate how different inputs affect the model. First, we feed RGB images and thermal (or depth) images into the network, respectively, following the unimodal crowd counting setting of the backbone network. Then we adopt the early fusion strategy, feeding the concatenation of RGB and thermal images into BL network. We also apply late fusion where the different modal features are extracted separately with the backbone network and then combine multimodal features in the last layer to estimate the crowd density map. From Table 5 and Table 6, we can observe an interesting finding that thermal images are more discriminative than RGB images for the RGBT-CC dataset. However, RGB images are more efficient than depth images on ShanghaiTechRGBD. Interestingly, naive feature fusion strategies such as early fusion and late fusion even damage the original accuracy.

To explore the effectiveness of each CSCA component, we remove the SCA block and CFA block, respectively. As shown in Table 5 and Table 6, regardless of which component is removed, the performance of our model would drop severely. These results suggest that the main components of CSCA work complementary. Additionally, we perform additional experiments replacing the SCA block with

Table 7. The performance of our model implemented with BL backbone under different illumination conditions on RGBT-CC.

Illumination	Input	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
	RGB	26.96	31.82	37.07	46.00	48.42
Brightness	T	19.19	24.07	30.28	39.93	32.07
	RGB-T	**14.41**	**18.85**	**24.71**	**34.20**	**24.74**
	RGB	45.78	51.03	54.65	62.51	87.29
Darkness	T	16.38	21.65	26.67	34.59	28.45
	RGB-T	**14.22**	**18.97**	**22.89**	**30.69**	**27.25**

the typical non-local block. Remarkably, our CSCA is far superior to non-local-based methods in terms of various evaluation metrics. Since the non-local-based methods were out of memory on the ShanghaiTechRGBD dataset where image resolution is 1920×1080, we could not report the results using non-local blocks. These results suggest that the SCA blocks have more practical potential than the typical non-local blocks.
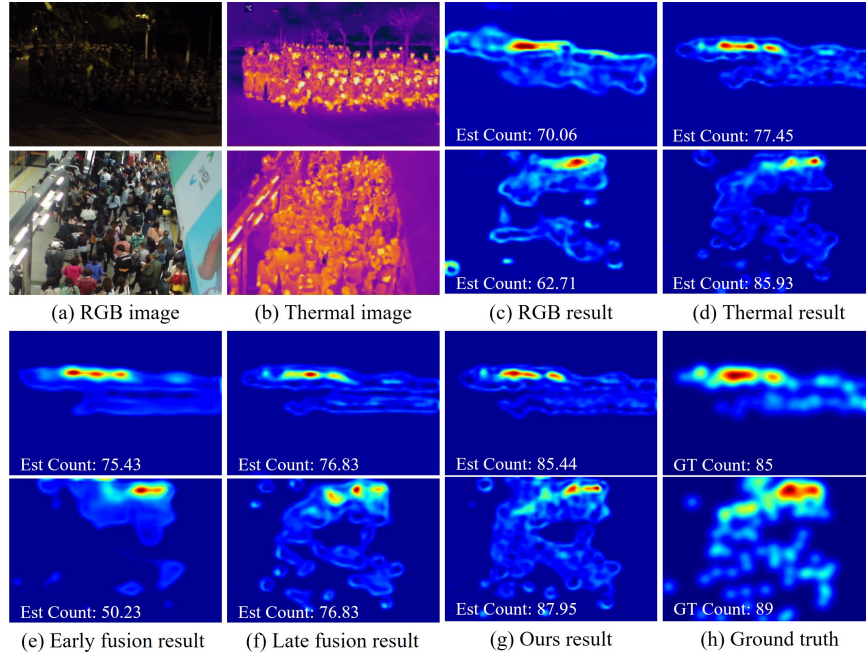
To further validate the complementarity of the multimodal data for crowd counting, we infer our model using BL backbone under different illumination conditions on the RGBT-CC dataset. As shown in Table 7, we conduct experiments under bright and dark scenes, with three different input settings, including uni-modal RGB images, unimodal thermal images, and RGB-T multimodal images, respectively. Notably, thermal images can provide more discriminative features than RGB images, especially in lower illumination cases. Moreover, comparing the single-modal results, it is easy to conclude that our CSCA can complementarily fuse thermal information and RGB optical information.

5   Qualitative Results

From the visualization results in cases (a) to (d) of Fig. 4, we can easily find that thermal images can facilitate the crowd counting task better than RGB images. Moreover, as shown in Fig. 4-(b), thermal images can also provide additional crowd location information to a certain extent, especially in dark scenes or bright scenes with less light noise. As we discussed earlier, inappropriate fusions fail to exploit the potential complementarity of multimodal data and even degrade the performance, such as the early fusion and late fusion shown in Fig. 4-(e) and Fig. 4-(f). Our proposed CSCA, a plug-and-play module, achieve significant improvements for cross-modal crowd counting by simply integrating into the backbone network as shown in Fig. 4-(g). This result shows the effectiveness of CSCA's complementary multi-modal fusion.

6   Conclusion

Whereas previous crowd counting approaches have mainly focused on RGB-based frameworks, we focus on a cross-modal representation that can addition-

(a) RGB image   (b) Thermal image   (c) RGB result   (d) Thermal result

Est Count: 70.06   Est Count: 77.45

Est Count: 62.71   Est Count: 85.93

Est Count: 75.43   Est Count: 76.83   Est Count: 85.44   GT Count: 85

Est Count: 50.23   Est Count: 76.83   Est Count: 87.95   GT Count: 89

(e) Early fusion result   (f) Late fusion result   (g) Ours result   (h) Ground truth

Fig. 4. Qualitative Results. (a) and (b) show the input RGB images and thermal images in different illumination conditions. (c)-(d) are the results of the RGB-based and thermal-based network. (e)-(f) are the results of the early fusion and late fusion. (g) and (h) refers to the results of our CSCA and the ground truth.

ally utilize thermal or depth images. Specifically, we propose a novel plug-and-play module named Cross-modal Spatio-Channel Attention (CSCA), which can be integrated with any RGB-based model architecture. Our CSCA blocks consist of two main modules. First, SCA is designed to spatially capture global feature correlations among multimodal data. The proposed SCA to deal with cross-modal relationships is computationally more efficient than existing non-local blocks by re-assembling the feature maps at the spatial level. Furthermore, our CFA module adaptively recalibrates the spatially-correlated features at the channel level. Extensive experiments on the RGBT-CC and ShanghaiTechRGBD datasets show the effectiveness and superiority of our method for multimodal crowd counting. Additionally, we conduct extensive comparative experiments in various scenarios using unimodal or multimodal representations, which have not been well addressed in crowd counting. These results are expected to be used as baselines for subsequent studies for cross-modal crowd counting.

References

1. Ali, A., Zhu, Y., Chen, Q., Yu, J., Cai, H.: Leveraging spatio-temporal patterns for predicting citywide traffic crowd flows using deep hybrid neural networks. In: IEEE Int'l Conf. on Parallel and Distributed Systems (ICPADS). pp. 125–132. IEEE (2019)
2. Bathija, A., Sharma, G.: Visual object detection and tracking using yolo and sort. International Journal of Engineering Research Technology **8**(11) (2019)
3. Bondi, E., Seidenari, L., Bagdanov, A.D., Del Bimbo, A.: Real-time people counting from depth imagery of crowded environments. In: Proc. of Int'l Conf. on Advanced Video and Signal Based Surveillance (AVSS). pp. 337–342. IEEE (2014)
4. Boominathan, L., Kruthiventi, S.S., Babu, R.V.: Crowdnet: A deep convolutional network for dense crowd counting. In: ACM Int'l Conf. on Multimedia (ACMMM). pp. 640–644 (2016)
5. Cao, X., Wang, Z., Zhao, Y., Su, F.: Scale aggregation network for accurate and efficient crowd counting. In: Proc. of European Conf. on Computer Vision (ECCV). pp. 734–750 (2018)
6. Chan, A.B., Liang, Z.S.J., Vasconcelos, N.: Privacy preserving crowd monitoring: Counting people without people models or tracking. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 1–7. IEEE (2008)
7. Chan, A.B., Vasconcelos, N.: Bayesian poisson regression for crowd counting. In: Proc. of Int'l Conf. on Computer Vision (ICCV). pp. 545–551. IEEE (2009)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. of Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. 886–893. Ieee (2005)
9. Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., Hu, Q., Peng, T., Zheng, J., Wang, X., Zhang, Y., et al.: Visdrone-det2019: The vision meets drone object detection in image challenge results. In: Proc. of Int'l Conf. on Computer Vision Workshops (ICCVW). pp. 0–0 (2019)
10. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: Survey and experiments. IEEE Trans. on Pattern Anal. Mach. Intell. (TPAMI) **31**(12), 2179–2195 (2008)
11. Fan, D.P., Zhai, Y., Borji, A., Yang, J., Shao, L.: Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In: Proc. of European Conf. on Computer Vision (ECCV). pp. 275–292. Springer (2020)
12. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. on Pattern Anal. Mach. Intell. (TPAMI) **32**(9), 1627–1645 (2010)
13. Fu, H., Ma, H., Xiao, H.: Real-time accurate crowd counting based on rgb-d information. In: IEEE Int'l Conf. on Image Processing (ICIP). pp. 2685–2688. IEEE (2012)
14. Fu, K., Fan, D.P., Ji, G.P., Zhao, Q.: Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 3052–3062 (2020)
15. Fu, M., Xu, P., Li, X., Liu, Q., Ye, M., Zhu, C.: Fast crowd density estimation with convolutional neural networks. Engineering Applications of Artificial Intelligence **43**, 81–88 (2015)
16. Ghodgaonkar, I., Chakraborty, S., Banna, V., Allcroft, S., Metwaly, M., Bordwell, F., Kimura, K., Zhao, X., Goel, A., Tung, C., et al.: Analyzing worldwide social distancing through large-scale computer vision. arXiv preprint arXiv:2008.12363 (2020)

17. Hong, S., Im, W., Yang, H.S.: Cbvmr: content-based video-music retrieval using soft intra-modal structure constraint. In: Int'l Conf. on Multimedia Retrieval (ICMR). pp. 353–361 (2018)
18. Hong, S., Kang, S., Cho, D.: Patch-level augmentation for object detection in aerial images. In: Proc. of Int'l Conf. on Computer Vision Workshops (ICCVW). pp. 0–0 (2019)
19. Idrees, H., Saleemi, I., Seibert, C., Shah, M.: Multi-source multi-scale counting in extremely dense crowd images. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 2547–2554 (2013)
20. Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M.: Composition loss for counting, density map estimation and localization in dense crowds. In: Proc. of European Conf. on Computer Vision (ECCV). pp. 532–546 (2018)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: Proc. of Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. 878–885. IEEE (2005)
23. Li, H., Zhang, S., Kong, W.: Rgb-d crowd counting with cross-modal cycle-attention fusion and fine-coarse supervision. IEEE Trans. on Industrial Informatics (2022)
24. Li, M., Zhang, Z., Huang, K., Tan, T.: Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: Proc. of Int'l Conf. on Pattern Recognition (ICPR). pp. 1–4. IEEE (2008)
25. Li, X., Hou, Y., Wang, P., Gao, Z., Xu, M., Li, W.: Trear: Transformer-based rgb-d egocentric action recognition. IEEE Trans. on Cognitive and Developmental Systems (2021)
26. Li, Y., Zhang, X., Chen, D.: Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 1091–1100 (2018)
27. Lian, D., Li, J., Zheng, J., Luo, W., Gao, S.: Density map regression guided detection network for rgb-d crowd counting and localization. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 1821–1830 (2019)
28. Lin, S.F., Chen, J.Y., Chao, H.X.: Estimation of number of people in crowded scenes using perspective transformation. IEEE Trans. on Systems, Man, and Cybernetics-Part A: Systems and Humans **31**(6), 645–654 (2001)
29. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proc. of Int'l Conf. on Computer Vision (ICCV). pp. 2980–2988 (2017)
30. Liu, J., Gao, C., Meng, D., Hauptmann, A.G.: Decidenet: Counting varying density crowds through attention guided detection and density estimation. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 5197–5206 (2018)
31. Liu, L., Chen, J., Wu, H., Li, G., Li, C., Lin, L.: Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 4823–4833 (2021)
32. Liu, X., Van De Weijer, J., Bagdanov, A.D.: Leveraging unlabeled data for crowd counting by learning to rank. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 7661–7669 (2018)
33. Liu, Z., He, Z., Wang, L., Wang, W., Yuan, Y., Zhang, D., Zhang, J., Zhu, P., Van Gool, L., Han, J., et al.: Visdrone-cc2021: The vision meets drone crowd

counting challenge results. In: Proc. of Int'l Conf. on Computer Vision (ICCV). pp. 2830–2838 (2021)

34. Ma, Z., Wei, X., Hong, X., Gong, Y.: Bayesian loss for crowd count estimation with point supervision. In: Proc. of Int'l Conf. on Computer Vision (ICCV). pp. 6142–6151 (2019)

35. Mehta, S., Rastegari, M.: Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 (2021)

36. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. Proc. of Neural Information Processing Systems (NeurIPS) **34**, 14200–14213 (2021)

37. Pang, Y., Zhang, L., Zhao, X., Lu, H.: Hierarchical dynamic filtering network for rgb-d salient object detection. In: Proc. of European Conf. on Computer Vision (ECCV). pp. 235–252. Springer (2020)

38. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Proc. of Neural Information Processing Systems (NeurIPS) **32** (2019)

39. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Proc. of Neural Information Processing Systems (NeurIPS) **28** (2015)

40. Shao, J., Loy, C.C., Kang, K., Wang, X.: Slicing convolutional neural network for crowd video understanding. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 5620–5628 (2016)

41. Shi, M., Yang, Z., Xu, C., Chen, Q.: Revisiting perspective information for efficient crowd counting. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 7279–7288 (2019)

42. Shi, Z., Mettes, P., Snoek, C.G.: Counting with focus for free. In: Proc. of Int'l Conf. on Computer Vision (ICCV). pp. 4200–4209 (2019)

43. Sindagi, V.A., Patel, V.M.: Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: Proc. of Int'l Conf. on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6. IEEE (2017)

44. Sindagi, V.A., Patel, V.M.: Generating high-quality crowd density maps using contextual pyramid cnns. In: Proc. of Int'l Conf. on Computer Vision (ICCV). pp. 1861–1870 (2017)

45. Sun, T., Di, Z., Che, P., Liu, C., Wang, Y.: Leveraging crowdsourced gps data for road extraction from aerial imagery. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 7509–7518 (2019)

46. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemannian manifolds. IEEE Trans. on Pattern Anal. Mach. Intell. (TPAMI) **30**(10), 1713–1727 (2008)

47. Valencia, I.J.C., Dadios, E.P., Fillone, A.M., Puno, J.C.V., Baldovino, R.G., Billones, R.K.C.: Vision-based crowd counting and social distancing monitoring using tiny-yolov4 and deepsort. In: 2021 IEEE International Smart Cities Conference (ISC2). pp. 1–7. IEEE (2021)

48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Proc. of Neural Information Processing Systems (NeurIPS) **30** (2017)

49. Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. Int'l Journal of Computer Vision (IJCV) **63**(2), 153–161 (2005)

50. Walach, E., Wolf, L.: Learning to count with cnn boosting. In: Proc. of European Conf. on Computer Vision (ECCV). pp. 660–676. Springer (2016)

51. Wan, J., Liu, Z., Chan, A.B.: A generalized loss function for crowd counting and localization. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 1974–1983 (2021)
52. Wang, C., Zhang, H., Yang, L., Liu, S., Cao, X.: Deep people counting in extremely dense crowds. In: ACM Int'l Conf. on Multimedia (ACMMM). pp. 1299–1302 (2015)
53. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 7794–7803 (2018)
54. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. Int'l Journal of Computer Vision (IJCV) **75**(2), 247–266 (2007)
55. Xu, M., Ge, Z., Jiang, X., Cui, G., Lv, P., Zhou, B., Xu, C.: Depth information guided crowd counting for complex crowd scenes. Pattern Recognition Letters **125**, 563–569 (2019)
56. Xu, M., Li, C., Lv, P., Lin, N., Hou, R., Zhou, B.: An efficient method of crowd aggregation computation in public areas. IEEE Trans. on Circuits and Systems for Video Technology. (TCSVT) **28**(10), 2814–2825 (2017)
57. Xu, T., Chen, W., Wang, P., Wang, F., Li, H., Jin, R.: Cdtrans: Cross-domain transformer for unsupervised domain adaptation. arXiv preprint arXiv:2109.06165 (2021)
58. Yang, Y., Li, G., Wu, Z., Su, L., Huang, Q., Sebe, N.: Weakly-supervised crowd counting learns from sorting rather than locations. In: Proc. of European Conf. on Computer Vision (ECCV). pp. 1–17. Springer (2020)
59. Zhang, A., Shen, J., Xiao, Z., Zhu, F., Zhen, X., Cao, X., Shao, L.: Relational attention network for crowd counting. In: Proc. of Int'l Conf. on Computer Vision (ICCV). pp. 6788–6797 (2019)
60. Zhang, H., Kyaw, Z., Chang, S.F., Chua, T.S.: Visual translation embedding network for visual relation detection. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 5532–5540 (2017)
61. Zhang, J., Fan, D.P., Dai, Y., Anwar, S., Saleh, F.S., Zhang, T., Barnes, N.: Ucnet: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 8582–8591 (2020)
62. Zhang, Q., Chan, A.B.: Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 8297–8306 (2019)
63. Zhang, S., Wu, G., Costeira, J.P., Moura, J.M.: Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In: Proc. of Int'l Conf. on Computer Vision (ICCV). pp. 3667–3676 (2017)
64. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 589–597 (2016)
65. Zhao, M., Zhang, J., Zhang, C., Zhang, W.: Leveraging heterogeneous auxiliary tasks to assist crowd counting. In: Proc. of Computer Vision and Pattern Recognition (CVPR). pp. 12736–12745 (2019)
66. Zhu, Z., Xu, M., Bai, S., Huang, T., Bai, X.: Asymmetric non-local neural networks for semantic segmentation. In: Proc. of Int'l Conf. on Computer Vision (ICCV). pp. 593–602 (2019)