MULTIMODAL CROWD COUNTING WITH MUTUAL ATTENTION TRANSFORMERS

Zhengtao Wu[1], *Lingbo Liu*[2], *Yang Zhang*[1], *Mingzhi Mao*[1,*], *Liang Lin*[1] *and Guanbin Li*[1,*]

[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Hong Kong Polytechnic University

ABSTRACT

Crowd counting is a fundamental yet challenging task that aims to automatically estimate the number of people in crowded scenes. Nowadays, with the rapid development of thermal and depth sensors, thermal images and depth maps become more accessible, which are proven to be beneficial information in boosting the performance of crowd counting. Consequently, we propose a Mutual Attention Transformer (MAT) module to fully leverage the complementary information of different modalities. Specifically, our MAT employs a cross-modal mutual attention mechanism to utilize the features of one modality to enhance the features of the other. Moreover, to improve performance by learning better visual representation and further exploiting modality-wise complementarity, we design a self-supervised pre-training method based on cross-modal image reconstruction. Extensive experiments on two standard benchmarks (i.e., RGBT-CC and ShanghaiTechRGBD) show that the proposed method is effective and universal for multimodal crowd counting, outperforming previous state-of-the-art methods.

Index Terms— Crowd Counting, Multimodal, Mutual Attention, Transformer, Self-Supervised Learning

1. INTRODUCTION

Crowd counting is a fundamental task in computer vision, whose purpose is to accurately and automatically count the number of pedestrians in images or surveillance videos. It has drawn increasing attention due to its wide range of practical applications, such as crowd control, urban planning, traffic management, etc. In the literature, numerous models have been proposed to address this task and have achieved considerable performance [1–7].

Nevertheless, most of the previous methods make predictions of crowd count only based on the optical information in RGB images, which may fail to perform accurate estimation in the wild when encountering poor illumination conditions or suffering severe variations of scale and perspective. The recent cross-modal approaches [8–10] show that incorporating thermal images or depth maps as additional information into RGB images delivers superior performance in crowd counting since thermal images or depth maps are highly complementary to RGB images. Specifically, as shown in Figure 1(a), thermal images are robust to illumination and can greatly help recognize possible pedestrians from cluttered backgrounds. Conversely, RGB images can help eliminate false positives in thermal images [8]. Likewise, as shown in Figure 1(b), the depth maps can exceedingly facilitate crowd counting by providing additional geometry information (e.g., size and location of heads) [9, 10], which alleviates the unfavorable effects of large scale variations and perspective changes. And in turn, the essential visual information of the crowd carried by RGB images can help remove the objects having similar appearances to pedestrians in depth maps. In a nutshell, additional information from other modalities (i.e., thermal and depth) and optical information (i.e., RGB images) are immensely complementary to each other.

(a) RGBT-CC [8]. (b) ShanghaiTechRGBD [9].

Fig. 1. Samples from RGBT-CC [8] dataset and ShanghaiTechRGBD [9] dataset. Images in the first and second row are the RGB images and thermal images (for RGBT-CC)/depth maps(for ShanghaiTechRGBD), respectively.

However, it remains challenging to explore and fuse the complementary information of different modalities. Conventional approaches [9–11] for multimodal data fusion either simply combine multimodal data before feeding them into networks or fuse features using ordinary fusion operations, both failing to fully capture the complementarity between modalities. Liu *et al.* [8] proposed an information aggregation and distribution module (IADM) with a shared branch to leverage modality-wise complementarity in crowd counting. However, IADM only uses local operations such as convolution, element-wise addition/subtraction, etc., which fails to

*Corresponding authors are Mingzhi Mao and Guanbin Li. This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No.2020B1515020048, in part by the National Natural Science Foundation of China under Grant No.61976250 and No.U1811463, and in part by the Guangzhou Science and technology project under Grant No.202102020633.

2022 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-8563-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICME52920.2022.9859777

model global cross-modal dependencies and relationships.

Recently, Transformers [12] have been widely applied to computer vision [13], since it was proven to have a strong potential to capture long-range dependencies by self-attention mechanism. Inspired by the success of Transformers, we propose a Mutual Attention Transformer (MAT) module to fully exploit the complementary information of different modalities. Specifically, our MATs are integrated into multiple stages of two modality-specific backbones to hierarchically learn cross-modal joint representation. At each stage, the features from each modality-specific backbone are converted into sequences of patch embeddings, and then we employ cross-modal multi-head mutual attention to dynamically propagate the patch embeddings of one modality to enhance the patch embeddings of the other. The enhanced patch embeddings of each modality are rearranged to image-shape features, which are further fed into the corresponding modality-specific backbone for higher-level representation learning. It is worth noting that our MAT is naturally effective for building cross-modal dependencies and relationships on both local and global levels, and therefore can fully capture the complementarities of different modalities for robust crowd counting.

For better visual representation and modality-wise complementarity learning, we design a novel and effective self-supervised pre-training method based on cross-modal image reconstruction to further boost performance. As a pretext task, the reconstruction-based method corrupts the inputs and learns to reconstruct them, during which the networks can learn feature representation of the data [14]. Based on this insight, we randomly mask (i.e., remove) blocks of the image pairs from two modalities and force the networks learn to reconstruct them by predicting the pixel values of the masked regions. The learned parameters are then transferred and fine-tuned on crowd counting. Most importantly, the key idea is that the reconstruction of some masked regions from one modality can be facilitated by the counterparts from the other, leading the networks to learn modality-wise complementarity.

In summary, our main contributions are three-fold:

- We propose Mutual Attention Transformer (MAT), a cross-modal fusion module, for multimodal crowd counting by fully leveraging the complementary information of different modalities.
- We develop a novel and effective self-supervised pre-training method based on cross-modal image reconstruction to further boost the performance of multimodal crowd counting.
- Extensive experiments on two challenging multimodal crowd counting benchmarks demonstrate the effectiveness and universality of our method, which achieves superior performance in comparison to previous state-of-the-art approaches.

2. RELATED WORKS

Crowd counting approaches: A large number of methods [1–7] with different network architectures were proposed for crowd counting. The mainstream methods are regression-based and usually generate density maps for crowd images and then sum up all the values of pixels to get the final counts. Meanwhile, various loss function designs [4, 15, 16] on crowd counting were put forward to improve performance and/or the quality of density maps. Recently, several works [8–10] introduced additional information from other modalities (i.e., thermal [8] or depth [9, 10]) to crowd counting for better performance.

Multimodal fusion methods: Multimodal fusion aims to properly integrate information from different modalities to make predictions for a specific task. Simple multimodal fusion approaches often combine features from different modalities using element-wise addition/multiplication or concatenation in the way of either "Early fusion" or "Late fusion" [9–11]. Besides, two-stream-based networks with hybrid fusion (i.e., hybrid of early and late fusion) [17–19] were proposed to hierarchically learn cross-modal features. However, most of these methods only consider the additional modality as an auxiliary one and adopt one-way information transfer. To leverage modality complementarities, a shared-branch-based approach was proposed [8]. Nevertheless, they cannot well capture the global long-range dependencies between modalities, which are vital information for crowd counting. In this paper, the proposed MAT module is naturally effective for dynamically building long-range dependencies on both local and global levels by utilizing mutual attention mechanism.

Transformers: Transformer [12], a new attention-based building block, was first proposed to tackle the machine translation tasks and successfully applied in natural language processing because of its strong ability to model global long-range dependencies. Recently, Transformers were extended to various computer vision tasks. For instance, ViT [13] converted an image into a sequence of flattened 2D patches and further processed it in a pure Transformer manner for image classification. DETR [20] adopted a hybrid of convolutional neural networks (CNN) and Transformers to address object detection. And VisTR [21] fed the features extracted by CNN of each video frame to a Transformer for video instance segmentation. Besides, some Transformer-based methods were proposed for multimodal tasks [22, 23]. Inspired by these works, we propose a Transformer-based cross-modal feature fusion method for multimodal crowd counting.

3. METHOD

In this work, we propose a multimodal crowd counting framework embedded with Mutual Attention Transformer (MAT) modules. Moreover, we design an image reconstruction-based cross-modal self-supervised pre-training method to fur-

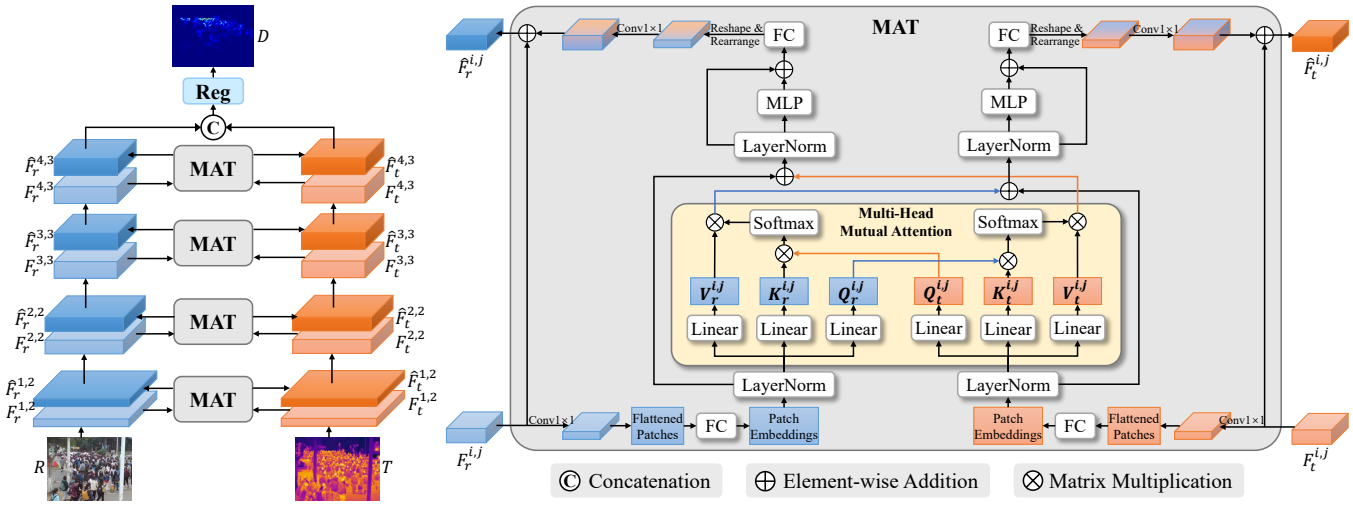

Fig. 2. The proposed framework for multimodal crowd counting. With the proposed Mutual Attention Transformer (MAT) modules embedded in several different layers between two modality-specific backbones, our method is capable of fully leveraging modality-wise complementarity. Note that "Reg" stands for regressor.

ther boost the performance of multimodal crowd counting. In this section, we utilize CSRNet [2] on RGBT-CC [8] dataset as an example to illustrate our method, where CSRNet is a representative crowd counting network and RGBT-CC is an RGB-thermal multimodal crowd counting dataset. Besides, to demonstrate our framework is effective and universal for different networks and datasets, we also implement it with the network BL [16] and conduct experiments on an RGB-depth dataset ShanghaiTechRGBD [9].

3.1. Overview

As shown in Figure 2, the proposed framework consists of two modality-specific backbones with several Mutual Attention Transformer (MAT) modules embedded in between, and a regressor ("Reg" module shown in Figure 2). Specifically, these two backbones hierarchically extract the features of RGB images and thermal images, respectively. Meanwhile, the MATs perform cross-modal feature fusion at several specific layers. Then, the concatenation of the last features of these two backbones is fed into the regressor to generate the final high-quality crowd density maps.

The aforementioned modality-specific backbones and regressor are implemented based on CSRNet. CSRNet consists of a front-end and a back-end, where the front-end is the first 10 convolutional layers of VGG16 [24] and the back-end is composed of six dilated convolutional layers along with a final 1×1 convolutional layer. In our framework, the modality-specific backbones are based on the front-end and the regressor is based on the back-end.

As shown in Figure 2, the network takes an RGB image R and a thermal image T as inputs and generates a crowd density map D. The two modality-specific backbones extract the features of R and T, respectively. For convenience, following Liu *et al.* [8], we denote the extracted features of R and T at layer Conv$i$_j as $F_r^{i,j}$ and $F_t^{i,j}$, respectively. To enable cross-modal feature fusion and fully leverage modality-wise complementarity hierarchically, we embed our MAT modules between the two modality-specific backbones at several different layers, which are Conv1_2, Conv2_2, Conv3_3, and Conv4_3. Specifically, the MAT embedded at layer Conv$i$_j takes $F_r^{i,j}$ and $F_t^{i,j}$ as inputs to perform cross-modal feature fusion by utilizing one to enhance the other. The operation of MAT can be formulated as follow:

$$\hat{F}_r^{i,j}, \hat{F}_t^{i,j} = \mathrm{MAT}(F_r^{i,j}, F_t^{i,j}), \tag{1}$$

where $\hat{F}_r^{i,j}$ and $\hat{F}_t^{i,j}$ are the enhanced features of $F_r^{i,j}$ and $F_t^{i,j}$, respectively. And then $\hat{F}_r^{i,j}$ and $\hat{F}_t^{i,j}$ are fed to the following layers of their respective backbones. Note that the last enhanced features (i.e., $\hat{F}_r^{4,3}$ and $\hat{F}_t^{4,3}$) are concatenated then fed to the regressor to generate the final density map D.

3.2. Mutual Attention Transformer

In order to fully exploit the modality-wise complementarity and capture global long-range dependencies between modalities, we propose a Mutual Attention Transformer (MAT) module. As depicted in Figure 2, MAT takes $F_r^{i,j}, F_t^{i,j}$ as inputs and outputs the corresponding enhanced features $\hat{F}_r^{i,j}, \hat{F}_t^{i,j}$ after feature fusion.

In general, the input features of a MAT are first converted to sequences of patch embeddings and then are fed to the multi-head mutual attention sub-module to perform cross-modal mutual attention and fusion by modeling long-range relationships of two modalities. After that, the fused patch embeddings are fed to two feed-forward networks, respectively, and then reshaped and rearranged to obtain the fused features. Finally, the enhanced features are obtained by com-

puting the element-wise addition of the fused features and the input features. For simplicity, in this section, we denote $F_r^{i,j}, F_t^{i,j}, \hat{F}_r^{i,j}$, and $\hat{F}_t^{i,j}$ as $F_r, F_t, \hat{F}_r$, and $\hat{F}_t$, respectively.

We first employ two 1×1 convolutional layers to reduce the number of channels of $F_r, F_t \in \mathbb{R}^{H\times W\times C}$, yielding the features $F_r', F_t' \in \mathbb{R}^{H\times W\times C'}$, respectively, where (H, W) is the spatial resolution of the features, C is the original number of channels, and C' is the reduced number of channels. The following multi-head mutual attention layer and the two feed-forward networks expect sequences as inputs, thus we reshape F_r', F_t' into two sequences of flattened 2D patches $\mathbf{x}_r^{(p)}, \mathbf{x}_t^{(p)} \in \mathbb{R}^{N\times d_{patch}}$, where $d_{patch} = P^2\cdot C'$ is the dimension of each flattened patches with the resolution of (P, P), and $N = HW/P^2$ is the number of patches. Then, we encode $\mathbf{x}_r^{(p)}, \mathbf{x}_t^{(p)}$ to patch embeddings $\mathbf{x}_r^{(emb)}, \mathbf{x}_t^{(emb)} \in \mathbb{R}^{N\times d}$ by two fully connected layers (FCs), respectively, where d is the dimension of each patch embeddings. Afterwards, $\mathbf{x}_r^{(emb)}, \mathbf{x}_t^{(emb)}$ are further processed by two layer normalization (LayerNorm) [25] layers, respectively.

As shown in Figure 2, the patch embeddings are then fed to the subsequent multi-head mutual attention sub-module. Specifically, $\mathbf{x}_r^{(emb)}$ and $\mathbf{x}_t^{(emb)}$ are linearly projected to produce their queries, keys, and values, respectively, which are denoted as $\boldsymbol{Q}_r, \boldsymbol{K}_r, \boldsymbol{V}_r \in \mathbb{R}^{N\times d}$ and $\boldsymbol{Q}_t, \boldsymbol{K}_t, \boldsymbol{V}_t \in \mathbb{R}^{N\times d}$, respectively. Following the vanilla Transformer [12], we employ Scaled Dot-Product Attention to perform mutual attention for each head, which can be formulated as:

$$\text{head}_{r,i} = \text{Softmax}(\boldsymbol{Q}_{r,i}\boldsymbol{K}_{t,i}^T/\sqrt{d_k})\boldsymbol{V}_{t,i}, \quad (2)$$

$$\text{head}_{t,i} = \text{Softmax}(\boldsymbol{Q}_{t,i}\boldsymbol{K}_{r,i}^T/\sqrt{d_k})\boldsymbol{V}_{r,i}, \quad (3)$$

where $\text{head}_{r,i}$ and $\text{head}_{t,i}$ are the i^{th} head for R and T, respectively, and $d_k = \frac{d}{h}$ where h is the number of heads. Then the outputs of each head are concatenated and fed to a series of operations including dropouts, residual connections, layer normalizations and feed-forward networks. After that, we obtain the fused patch embeddings $\dot{\mathbf{x}}_r^{(emb)}, \dot{\mathbf{x}}_t^{(emb)} \in \mathbb{R}^{N\times d}$.

Next, we utilize two FCs to decode $\dot{\mathbf{x}}_r^{(emb)}, \dot{\mathbf{x}}_t^{(emb)}$ to two sequences of fused flatten patches $\dot{\mathbf{x}}_r^{(p)}, \dot{\mathbf{x}}_t^{(p)} \in \mathbb{R}^{N\times d_{patch}}$, respectively, which are then reshaped and rearranged to obtain the fused features with C' channels, respectively. Afterwards, we use another two 1×1 convolutional layers to recover the number of channels, generating the fused features $\dot{F}_r, \dot{F}_t \in \mathbb{R}^{H\times W\times C}$, respectively.

Finally, we compute element-wise addition of the fused features and the input features to get the enhanced features, which can be expressed as:

$$\hat{F}_r = F_r + \dot{F}_r, \quad \hat{F}_t = F_t + \dot{F}_t. \quad (4)$$

3.3. Cross-Modal Self-Supervised Pre-Training Method

As discussed in previous sections, our MATs excel at modeling long-range dependencies and relationships across different modalities to exploit modality-wise complementarity.

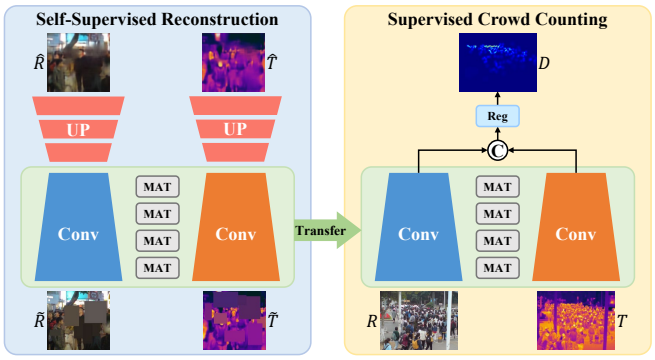

Fig. 3. The proposed cross-modal self-supervised pre-training method for multimodal crowd counting.

For better visual representation and modality-wise complementarity learning, we design an image reconstruction-based cross-modal self-supervised pre-training method to further boost performance for multimodal crowd counting.

As shown in Figure 3, the network for self-supervised reconstruction takes a pair of masked RGB image $\tilde{R}$ and masked thermal image $\tilde{T}$ as inputs and output their reconstructed images $\hat{R}$ and $\hat{T}$, respectively. Specifically, the procedure of masking an image is: (1) choose a random rectangle box in the image; (2) fill the box with its pixels' mean value; (3) repeat step (1)-(2) for n times, where $n \sim \mathcal{U}(n_{min}, n_{max})$ and $\mathcal{U}$ stands for uniform distribution. Specifically, the network consists of two modality-specific backbones with MATs embedded in between, which are the same as the ones described in Section 3.2, and two reconstructors. Each of the reconstructors is composed of three "Up" modules and a 1×1 convolutional layer, where an "Up" module contains a deconvolutional layer and two convolutional layers. After pre-training, the reconstructors are removed. The remaining parts of the network, having learned good visual representation and modality-wise complementarity, are then transferred and fine-tuned on crowd counting.

The network learns to reconstruct the original images R, T by predicting the pixel values of the masked regions. Standard pixel-wise reconstruction loss is computed in the masked regions, which is defined as follows:

$$\mathcal{L}_r = \frac{1}{N}\sum_{i=1}^{N}\sum_{p\in M_i}(\|\hat{R}_i(p) - R_i(p)\| + \|\hat{T}_i(p) - T_i(p)\|), \quad (5)$$

where N is the number of training samples, M_i is the union region of all masked regions of both $\tilde{R}_i$ and $\tilde{T}_i$, and p is the index of a pixel. Note that we do not use extra training data during pre-training.

4. EXPERIMENTS

4.1. Implementation Details and Evaluation Metrics

In this work, we implement our multimodal crowd counting framework based on CSRNet [2] and BL [16] on a single

NVIDIA RTX 2080Ti GPU. We reduce 30% of the channels of every convolutional layer in the backbones to keep the number of parameters similar to that of the original models for fair comparisons. Geometry-adaptive Gaussian kernels [1] are applied to generate the ground truth density maps. Adam [26] is adopted to optimize our networks. The learning rate is set to 1e-5 for crowd counting and 1e-4 for self-supervised pre-training. The filters' weights are randomly initialized by Gaussian distributions with zero mean and standard deviation of 0.01.

Root Mean Square Error (RMSE) and Grid Average Mean Absolute Error (GAME) [27] are used for evaluation. Specifically, for GAME at level L, we divide an image into 4^L non-overlapping regions and calculate the counting error in each region, which is defined as:

$$\mathrm{GAME}(L) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{4^L}\left|\hat{y}_i^j - y_i^j\right|, \tag{6}$$

where N is the number of test samples, $\hat{y}_i^j$ and y_i^j are the estimated count and the ground truth count in the j^{th} region of the i^{th} image, respectively. In particular, GAME(0) is equivalent to Mean Absolute Error (MAE).

4.2. Datasets

RGBT-CC: The recently proposed RGBT-CC [8] is a large-scale free-view multimodal crowd counting benchmark and contains 2,030 RGB-thermal image pairs. The number of image pairs used for training, validation, and testing is 1,030, 200, and 800, respectively. It is an extremely challenging dataset whose images are captured under different illumination conditions from various scenes, such as malls, streets, playgrounds, stations, etc., which has an average of 68 pedestrians marked with point annotations per image.

ShanghaiTechRGBD: ShanghaiTechRGBD [9] is a large-scale RGB-depth crowd counting dataset, consisting of 2,193 samples captured by surveillance cameras where each of which contains an RGB image and a depth map. The dataset has an average of 65.9 annotated heads per image. Specifically, 1,193 samples in ShanghaiTechRGBD are used for training and the remaining are for testing.

4.3. Comparison with State-of-the-art Methods

In this section, we evaluate and compare our proposed method implemented based on CSRNet and BL with other state-of-the-art methods on RGBT-CC and ShanghaiTechRGBD.

As shown in Table 1, both of our CSRNet-based and BL-based methods outperform their corresponding IADM-based [8] methods on all evaluation metrics by a large margin. For example, on RGBT-CC, compared to CSRNet+IADM, our CSRNet-based method (i.e., CSRNet+Ours) achieves 23.9% and 27.1% lower error of GAME(0) and RMSE, respectively. And BL+Ours achieves an improvement of 20.9% and 20.1% on GAME(0) and RMSE than BL+IADM [8], respectively. In particular, BL+Ours becomes the new state-of-the-art method on RGBT-CC. In addition, on ShanghaiTechRGBD, CSRNet+Ours surpasses the existing best method (i.e., DPDNet [10]) on all evaluation metrics. It is worth noting that CSRNet+Ours achieves an improvement of 16.3% and 21.8% on GAME(0) and RMSE over DPDNet, respectively, and becomes the new state-of-the-art method on ShanghaiTechRGBD. The outstanding performance well demonstrates the superiority of our method. Thanks to our tailor-designed MAT modules and self-supervised pre-training scheme, our method can fully capture the modality-wise complementarity by modeling dependencies and relationships across different modalities on both local and global levels, resulting in excellent performance.

Table 1. Performance of different methods on RGBT-CC and ShanghaiTechRGBD.

Method	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
RGBT-CC					
UCNet [17]	33.96	42.42	53.06	65.07	56.31
HDFNet [18]	22.36	27.79	33.68	42.48	33.93
BBSNet [19]	19.56	25.07	31.25	39.24	32.48
MVMS [28]	19.97	25.10	31.02	38.91	33.97
CSRNet+IADM [8]	17.94	21.44	26.17	33.33	30.91
CSRNet+Ours	**13.65**	**18.03**	**22.94**	**28.65**	**22.53**
BL+IADM [8]	15.61	19.95	24.69	32.89	28.18
BL+Ours	**12.35**	**16.29**	**20.81**	**29.09**	**22.53**
ShanghaiTechRGBD					
UCNet [17]	10.81	15.24	22.04	32.98	15.70
HDFNet [18]	8.32	13.93	17.97	22.62	13.01
BBSNet [19]	6.26	8.53	11.80	16.46	9.26
DetNet [29]	9.74	-	-	-	13.14
CL [15]	7.32	-	-	-	10.48
RDNet [9]	4.96	-	-	-	7.22
DPDNet [10]	4.23	5.67	7.04	9.64	6.75
BL+IADM [8]	7.13	9.28	13.00	19.53	10.27
BL+Ours	**5.39**	**6.73**	**8.98**	**13.66**	**7.77**
CSRNet+IADM [8]	4.38	5.95	8.02	11.02	7.06
CSRNet+Ours	**3.54**	**4.82**	**6.52**	**9.07**	**5.28**

Table 2. Ablation studies. "+SSP" indicates the model is self-supervised pre-trained using our method described in Section 3.3 then fine-tuned on crowd counting.

Method	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
RGBT-CC					
CSRNet (Early fusion)	20.40	23.58	28.03	35.51	35.26
CSRNet (Late fusion)	19.87	25.60	31.93	41.60	35.09
CSRNet+MAT	**15.39**	**19.70**	**24.65**	**30.75**	**25.55**
CSRNet+MAT+SSP	**13.65**	**18.03**	**22.94**	**28.65**	**22.53**
BL (Early fusion)	18.70	22.55	26.83	34.62	32.67
BL (Late fusion)	17.18	21.07	25.75	33.72	33.32
BL+MAT	**13.61**	**18.08**	**22.79**	**31.35**	**24.48**
BL+MAT+SSP	**12.35**	**16.29**	**20.81**	**29.09**	**22.53**
ShanghaiTechRGBD					
BL (Early fusion)	8.94	11.57	15.68	22.49	12.49
BL (Late fusion)	8.23	9.91	13.12	19.59	12.44
BL+MAT	**6.61**	**8.16**	**10.90**	**16.62**	**9.39**
BL+MAT+SSP	**5.39**	**6.73**	**8.98**	**13.66**	**7.77**
CSRNet (Early fusion)	4.92	6.78	9.47	13.06	7.41
CSRNet (Late fusion)	4.94	6.54	8.75	11.96	7.11
CSRNet+MAT	**4.05**	**5.49**	**7.49**	**10.25**	**6.01**
CSRNet+MAT+SSP	**3.54**	**4.82**	**6.52**	**9.07**	**5.28**

4.4. Ablation Studies

To convincingly demonstrate the effectiveness of each component of our method, we conduct extensive ablation studies

on both RGBT-CC and ShanghaiTechRGBD as shown in Table 2. "CSRNet (Early fusion)" method feeds the concatenation of RGB and thermal/depth images into the vanilla CSRNet. "CSRNet (Late fusion)" concatenates the features of RGB and thermal/depth images extracted by two separated backbones to generate density maps. Note that "CSRNet (Late fusion)" is equivalent to our proposed framework with all MATs removed. Similar rules are applied on BL.

Effectiveness of MAT: When equipped with MATs, we can observe that both CSRNet and BL achieve significantly superior performance on RGBT-CC and ShanghaiTechRGBD. For instance, on RGBT-CC, RMSE of CSRNet+MAT and RMSE of BL+MAT are remarkably reduced by 27.2% and 26.5%, respectively, compared to their "Late fusion" methods.

Effectiveness of self-supervised pre-training (SSP): The proposed self-supervised pre-training method can further boost performance considerably. For example, compared to CSRNet+MAT and BL+MAT on ShanghaiTechRGBD, CSRNet+MAT+SSP and BL+MAT+SSP further enjoy 12.2% and 17.3% of improvement on RMSE, respectively.

5. CONCLUSION

In this work, we propose a Mutual Attention Transformer (MAT) module for multimodal crowd counting. Our MAT can fully exploit the modality-wise complementarity to perform feature fusion by modeling cross-modal dependencies and relationships on both local and global levels. In addition, we develop a novel and effective self-supervised pre-training method based on cross-modal image reconstruction to further boost performance. Extensive experiments on two challenging multimodal crowd counting benchmarks demonstrate the effectiveness and universality of our method.

6. REFERENCES

[1] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma, "Single-image crowd counting via multi-column convolutional neural network," in *CVPR*, 2016, pp. 589–597.

[2] Yuhong Li, Xiaofan Zhang, and Deming Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *CVPR*, 2018, pp. 1091–1100.

[3] Lingbo Liu, Hongjun Wang, Guanbin Li, Wanli Ouyang, and Liang Lin, "Crowd counting using deep recurrent spatial-aware network," in *IJCAI*, 2018, pp. 849–855.

[4] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin, "Crowd counting with deep structured scale integration network," in *ICCV*, 2019, pp. 1774–1783.

[5] Zhilin Qiu, Lingbo Liu, Guanbin Li, Qing Wang, Nong Xiao, and Liang Lin, "Crowd counting via multi-view scale aggregation networks," in *ICME*, 2019, pp. 1498–1503.

[6] Lixian Yuan, Zhilin Qiu, Lingbo Liu, Hefeng Wu, Tianshui Chen, Pei Chen, and Liang Lin, "Crowd counting via scale-communicative aggregation networks," *Neurocomputing*, vol. 409, pp. 420–430, 2020.

[7] Lingbo Liu, Jiaqi Chen, Hefeng Wu, Tianshui Chen, Guanbin Li, and Liang Lin, "Efficient crowd counting via structured knowledge transfer," in *ACM MM*, 2020, pp. 2645–2654.

[8] Lingbo Liu, Jiaqi Chen, Hefeng Wu, Guanbin Li, Chenglong Li, and Liang Lin, "Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting," in *CVPR*, 2021, pp. 4823–4833.

[9] Dongze Lian, Jing Li, Jia Zheng, Weixin Luo, and Shenghua Gao, "Density map regression guided detection network for rgb-d crowd counting and localization," in *CVPR*, 2019, pp. 1821–1830.

[10] Dongze Lian, Xianing Chen, Jing Li, Weixin Luo, and Shenghua Gao, "Locating and counting heads in crowds with a depth prior," *TPAMI*, pp. 1–1, 2021.

[11] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao, "Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection," in *CVPR*, 2020, pp. 3052–3062.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017, vol. 30.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[14] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016, pp. 2536–2544.

[15] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *ECCV*, 2018, pp. 532–546.

[16] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong, "Bayesian loss for crowd count estimation with point supervision," in *ICCV*, 2019, pp. 6142–6151.

[17] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes, "Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders," in *CVPR*, 2020, pp. 8582–8591.

[18] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu, "Hierarchical dynamic filtering network for rgb-d salient object detection," in *ECCV*, 2020, pp. 235–252.

[19] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao, "Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network," in *ECCV*, 2020, pp. 275–292.

[20] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020, pp. 213–229.

[21] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia, "End-to-end video instance segmentation with transformers," in *CVPR*, 2021, pp. 8741–8750.

[22] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han, "Visual saliency transformer," in *ICCV*, 2021, pp. 4722–4732.

[23] Ronghang Hu and Amanpreet Singh, "Unit: Multimodal multitask learning with a unified transformer," in *ICCV*, 2021, pp. 1439–1449.

[24] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[25] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[26] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[27] Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto López-Sastre, Saturnino Maldonado-Bascón, and Daniel Onoro-Rubio, "Extremely overlapping vehicle counting," in *IbPRIA*, 2015, pp. 423–431.

[28] Qi Zhang and Antoni B. Chan, "Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns," in *CVPR*, 2019, pp. 8289–8298.

[29] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G. Hauptmann, "Decidenet: Counting varying density crowds through attention guided detection and density estimation," in *CVPR*, 2018, pp. 5197–5206.