RGB-T Multi-Modal Crowd Counting Based on Transformer

Zhengyi Liu
liuzywen@ahu.edu.cn
Wei Wu
2640947588@qq.com
Yacheng Tan
1084043983@qq.com
Guanghui Zhang
2532950974@qq.com

School of Computer Science and Technology
Anhui University
Hefei, China

Abstract

Crowd counting aims to estimate the number of persons in a scene. Most state-of-the-art crowd counting methods based on color images can't work well in poor illumination conditions due to invisible objects. With the widespread use of infrared cameras, crowd counting based on color and thermal images is studied. Existing methods only achieve multi-modal fusion without count objective constraint. To better excavate multi-modal information, we use count-guided multi-modal fusion and modal-guided count enhancement to achieve the impressive performance. The proposed count-guided multi-modal fusion module utilizes a multi-scale token transformer to interact two-modal information under the guidance of count information and perceive different scales from the token perspective. The proposed modal-guided count enhancement module employs multi-scale deformable transformer decoder structure to enhance one modality feature and count information by the other modality. Experiment in public RGBT-CC dataset shows that our method refreshes the state-of-the-art results. https://github.com/liuzywen/RGBTCC

1 Introduction

Crowd counting can predict the distribution of crowd and estimate the number of persons in unconstraint scenes. It is widely studied by the academia and industrial communities since the number of persons is an important indicator of incident monitoring[31], traffic control[19], and infectious disease prevention[32]. The existing crowd counting methods have achieved tremendous improvement due to the introduce of convolutional neural networks [7, 8] and transformer[28, 40].

However, when light is insufficient, the performance of crowd counting is unsatisfying, as shown in the first line of Fig.1. The thermal image can percept the temperature of objects to recognize the persons. Therefore, RGB-Thermal (RGB-T) crowd counting by introducing the thermal modality has attracted a lot of attentions.

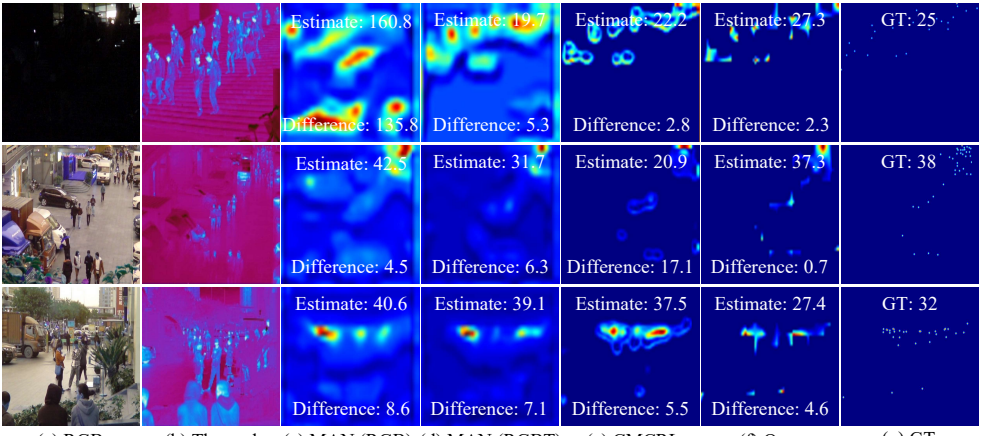

Figure 1: Three examples to show the performances of different methods in poor light condition, thermal disturbance, and large-scale variation, respectively. “Estimate" means predicted counts. “Difference" means the counting difference from the ground truth. (a) RGB image (b) the paired thermal image (c) MAN[18] result based on RGB image (d) MAN[18] result based on RGB-T image (e) CMCRL[20] result (f) our result (g) ground truth.

In RGB-T crowd counting task, a most important challenge is the multi-modal fusion problem. The color modality is good at perceiving the shape and texture of persons, but it is also interfered by the cluttered background. The thermal modality is skilled in recognizing persons which have temperature from scattered environments, but it also highlights the other heating objects, as shown in the second line of Fig.1 where the cars in the right are highlighted. Existing RGB-T crowd counting methods fuse complementary multi-modal features by Information Aggregation and Distribution Module (IADM) [20], Information Improvement Module (IIM) [29], and Mutual Attention Transformer (MAT) [41]. However, these multi-modal interactions are lack of a constraint. If we add a counting constraint on multi-modal fusion process, two-modal fusion has a clear goal. Therefore, we use a transformer structure to fuse two-modal information and design a learnable count token to participant the two-modal fusion. It makes the color and thermal modality interact under the guidance of a common count token.

In RGB-T crowd counting task, the other challenge is large-scale variation which is also the common issue in crowd counting, as shown in the third line of Fig.1 where persons that are far from the camera appear much smaller than those close to it. Existing methods use multi-column structure [1, 21, 46, 48], dilated convolution [2, 6, 15], high-resolution representation [24], and attention mechanism [18] to enlarge the receptive fields. Under the transformer framework, we propose a multi-scale token transformer to perceive persons with different scales. The tokens are merged to form token sequences with different lengths and then fed into some parallel transformers. After the enhancement of transformers, the receptive fields of features will be diversified.

To further improve the accuracy of crowd counting, we use a modality to guide the learning of the other modality and count token. A multi-scale deformable transformer is adopted to decode a modality and count token by the other modality. As a result, the count ability of the feature is enhanced.

In summary, the main contributions are summarized as follows:

- An RGB-T multi-modal crowd counting model is proposed based on the transformer. Multi-head self-attention is used to achieve the count-guided multi-modal fusion. Multi-head cross-attention is adopted to achieve the modal-guided count enhancement.
- A count-guided multi-modal fusion transformer is proposed to solve the fusion problem. Under the guidance of count global information, color and thermal modalities are well combined and aligned.
- A multi-scale token transformer is proposed to solve the large-scale variation problem. Three-scale token sequences are parallel handled to achieve multi-scale concept.
- The ablation experiments verify the effectiveness of modules, multi-scale design, and count guidance. The comparison experiments show the significant improvement over existing RGB-T crowd counting methods.

2 Related work

2.1 Crowd counting

Crowd counting can be achieved by detection [12, 13, 16, 27] or density map estimation [3, 14, 36, 38]. Since the latter can solve high overlap and occlusion problem, it shows better performance than the former.

The large-scale variation generated by the wide viewing angle of cameras and 2D perspective projection is a major challenge in crowd counting. The persons which are close to the camera are large, while the persons which are far from the camera are small. Multi-scale architecture[1, 2, 6, 15, 47, 48] and perspective information[9, 25, 42, 44, 45] are two main solutions. Recently, to deal with the scale changes, some attention based methods are proposed. MAN[18] improves global attention in the transformer by adding region attention. HANet[35] introduces scale context in the parallel spatial attention and channel attention.

In the paper, we solve the large-scale variation problem by multi-scale transformer based on tokens. The original token sequence is merged into a middle-scale token sequence and a large-scale token sequence, respectively. Then the three are parallel handled by three multi-head self-attention structures. Finally, three branches are concatenated and combined. The multi-scale concept ensures abundant receptive fields which benefits the crowd counting task.

2.2 Transformer based crowd counting

Previous works utilize the convolution neural network as the backbone and regress density map to predict the crowd count. The advent of transformer has pushed the crowd counting model forward. BCCTrans [28] introduces a global context learnable token to guide the counting. SAANet [40] designs a deformer backbone to extract the features, aggregates multi-level features by a deformable transformer encoder, and introduces a count query in a transformer decoder to re-calibrates the multi-level feature maps. DCSwinTrans[10] enhances the large-range contextual information by a dilated Swin Transformer backbone, and equips with a feature pyramid networks decoder to achieve crowd instant localization. CrowdFormer [43] models the human top-down visual perception mechanism by an overlap

patching transformer block. CCTrans [30] adopts a pyramid transformer and a multi-scale regression head to achieve both fully-supervised and weakly-supervised crowd counting task. In addition, in weakly-supervised crowd counting, there are some other transformer based methods. TransCrowd [17] uses a learnable counting token or global average pooling on high-layer semantic tokens to represent the crowd numbers. It constructs a weakly supervised model from sequence-to-count perspective. SFSL [5] introduces a learnable unbiased feature estimation of persons and utilizes the feature similarity for the regression of crowd numbers to solve the lack of local supervision. CrowdMLP [37] proposes a multi-granularity multilayer perceptron (MLP) regressor to enlarge receptive fields and a split-counting to decouple spatial constraints. JCTNet [34] introduces transformer structure upon the high-layer feature of convolutional neural network and regresses the count.

In the paper, we use transformer encoder structure to achieve count-guided multi-modal fusion, and use transformer decoder structure to perform modal-guided count enhancement.

2.3 RGB-T crowd counting

Although the crowd counting methods have achieved many significant improvements, they rely on optical information and often perform poorly when the light is insufficient. To solve this problem, RGB-T crowd counting has been getting a lot of attentions. On one hand, thermal image can recognize pedestrians in poor illumination conditions. On the other hand, thermal image can reduce wrong recognition about some human-shaped objects. Meanwhile, RGB image can suppress interference in thermal images. For example, heating walls and lamps that are highlighted in thermal images can be filtered from color perspective. Therefore, RGB and thermal images need to be simultaneously explored.

CMCRL [20] introduces a two-stream framework that first aggregates two features and second propagates the common information to further refine each feature. TAFNet [29] uses a three-stream network to learn the RGB feature, the thermal feature, and the concatenated RGB-T feature for crowd counting. The proposed Information Improvement Module (IIM) is used to fuse the modal-specific and combination features. Mutual Attention Transformer (MAT) [41] uses cross-modal mutual attention to build long-range dependencies and enhance semantic features in crowd counting task. DEFNet [49] uses multi-modal fusion, receptive field enhancement, and multi-layer fusion to highlight the crowd position and suppress the background noise.

In these works, the fusion of the RGB and thermal images are short of count objective constraint. We design a learnable count token to guide multi-modal fusion.

3 Proposed Method

We propose an RGB-T multi-modal crowd counting method which includes a count-guide multi-modal fusion, a modal-guide count enhancement, and a regression head, as shown in Fig.2. To solve multi-modal fusion problem, we introduce a count guidance. Moreover, to perceive the large-scale variation, we propose a multi-scale token concept. Combining both, multi-modal features are well fused towards a global objective. Furthermore, counting information is further enhanced from one modality under the guidance of the other modality.

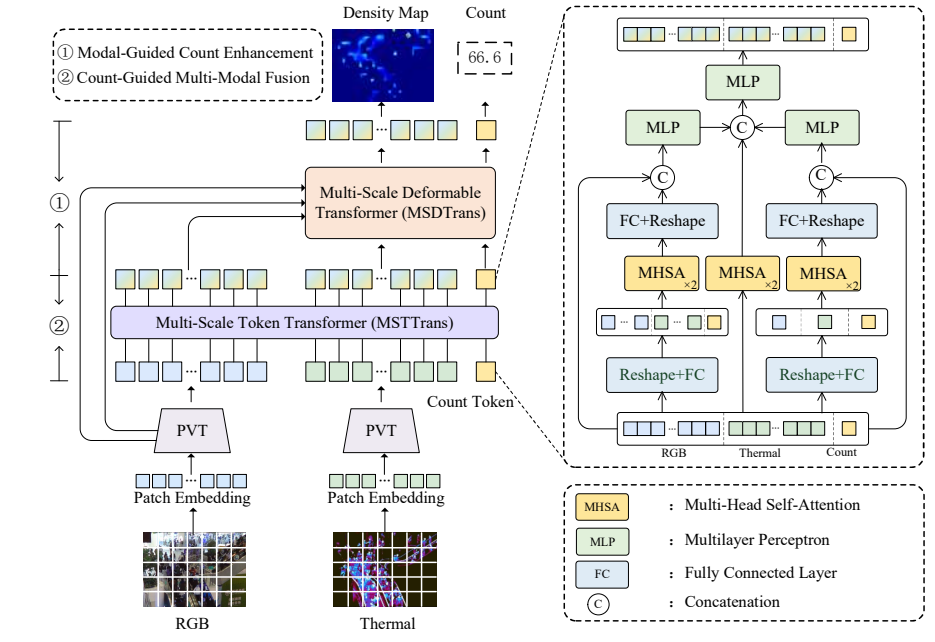

Figure 2: Our proposed RGB-T multi-modal crowd counting model based on transformer.

3.1 Count guided multi-modal fusion

Given a paired RGB-T image $I = \{I_r, I_t\}$, we use two PVT encoders [39] as the feature extractors to capture hierarchical features.

$$\begin{aligned} F_r &= \mathcal{E}_{PVT}(I_r) \\ F_t &= \mathcal{E}_{PVT}(I_t) \end{aligned} \tag{1}$$

where $\mathcal{E}_{PVT}$ denotes a PVT encoder, $F_r = \{F_r^i\}_{i=1}^4$ and $F_t = \{F_t^i\}_{i=1}^4$ represent color features and thermal features, respectively, i is the feature layer number.

The high-layer features contain more semantic information, which are suitable to obtain the global counting cues. Besides, color feature and thermal feature have each advantage in representing the crowd. Therefore, we use the high-level tokens from color modality and thermal modality to excavate the number of crowd. To fully align two-modal data and generate a consistent result, a learnable count token is designed to guide the two-modal fusion. Specifically, as is illustrated in Fig.2, high-layer semantic features F_r^4 and F_t^4 are generated from color encoder and thermal encoder, respectively. They represent unaligned multi-modal semantic concept. We design a learnable count token F_{count} which implies the coarse number of crowd. The three are concatenated along the token direction, and then fed into a Multi-Scale Token Transformer (*MSTTrans*) which spreads information among color, thermal, and crowd count by the multi-head self-attention.

MSTTrans is proposed to solve large-scale variations. Inspired by multi-scale design in atrous spatial pyramid pooling (ASPP) [4], *MSTTrans* achieves multi-scale transformer based on tokens. At first, we concatenate high-layer color feature, high-layer thermal feature, and the learnable count token to form an initial token sequence. Then, we merge the initial

token sequence to generate a middle-scale token sequence. The middle-scale token sequence has the larger receptive fields than original token sequence. Besides, we merge the initial token sequence to generate a large-scale token sequence, where a modality is represented by a token. According to above merge strategy, three parallel branches which all include color modality, thermal modality, and count token are constructed. They are fed into three multi-head self-attention modules for in-depth fusion.

Specifically, as is illustrated in the right of Fig. 2, suppose the high-layer semantic feature $F_r^4 \in \mathbb{R}^{N^2 \times C}$ and $F_t^4 \in \mathbb{R}^{N^2 \times C}$, where N^2 and C represent the number of tokens and channels, respectively. The two-modal features and the learnable count token are concatenated to generate the initial token sequence $f_1 \in \mathbb{R}^{(2N^2+1) \times C}$.

$$f_1 = [F_r^4, F_t^4, F_{count}] \tag{2}$$

where $[\cdot]$ denotes concatenation operation along token direction.

Then, the two-modal features are merged to N groups and each group generates N middle-scale tokens. All the middle-scale tokens and the learnable count token are concatenated to generate the middle-scale token sequence $f_2 \in \mathbb{R}^{(2N+1) \times C}$.

$$f_2 = [merge_{N^2 \to N}(F_r^4), merge_{N^2 \to N}(F_t^4), F_{count}] \tag{3}$$

where $merge_{a \to b}$ denotes the aggregation operation from a tokens to b tokens which applies a reshape operation and a fully connected layer.

Meanwhile, the two-modal features are merged to two groups and each group generates a large-scale token. The large-scale tokens and the learnable count token are concatenated to generate the large-scale token sequence $f_3 \in \mathbb{R}^{(2+1) \times C}$. There are a color token, a thermal token, and a learnable count token. It ensures two-modal whole alignment under the guidance of count.

$$f_3 = [merge_{N^2 \to 1}(F_r^4), merge_{N^2 \to 1}(F_t^4), F_{count}] \tag{4}$$

Three token sequences with different scales are fed into three multi-head self-attention modules for multi-modal interaction.

$$f_i' = MHSA(f_i) \tag{5}$$

where *MHSA* represents two multi-head self-attention layers.

Since the lengths of middle-scale and large-scale token sequences are different from initial token sequences, we apply fully connection layer and reshape operation to restore token sequence length.

$$g_i = Reshape(FC(f_i')) \tag{6}$$

where $i = 2, 3$ because only middle-scale and large-scale token sequences should be restored, *FC* is a fully-connected layer, and *Reshape* is a reshape operation to restore token length.

Further, to retain the original features in the middle-scale and large-scale branches, the concatenation and *MLP* operations are successively conducted.

$$g_i' = MLP(Concat(g_i, f_1)) \tag{7}$$

where $i = 2, 3$, *Concat* is concatenation operation along channel direction, and *MLP* is a two-layer perceptron.

Last, three features are concatenated and shrunk in channels.

$$G = [G_r, G_t, G_{count}] = MLP(Concat(f_1', g_2', g_3')) \tag{8}$$

where G has the same size as the input f_1 of *MSTTrans* module, and consists of optimized color feature G_r, thermal feature G_t, and count feature G_{count}.

In *MSTTrans* module, the count token is responsible for incorporating the global information and perceiving the number of persons. Besides, it is used to guide the fusion of color feature and thermal feature. Under the guidance of count token, color feature and thermal feature are in-depth interacted. Moreover, multi-scale token concept ensures the abundant receptive fields adaptive to recognizing the persons with different sizes.

3.2 Modal-guided counting enhancement

The researches pointed out that the thermal image can provide strong support on density map estimation, especially in the dark background[29]. In the paper, we use the thermal modality to predict the density map and count, and further use color modality to refine the prediction.

Therefore, after the previous count-guided multi-modal fusion, we design a modal-guided counting enhancement module which is responsible for generating the density map and final count from one modality under the guidance of the other modality. A multi-scale deformable transformer (*MSDTrans*) is employed to achieve the above objective.

Specifically, the thermal feature G_t and the learnable count token G_{count} are concatenated as query (Q), and the enhanced color feature G_r and the encoded low-layer features $F_r^i (i = 1,2,3)$ compose multi-scale color features which are regarded as key (K) and value (V). We use multi-scale deformable attention [50] to enhance Q by K and V. Last, it will output modal-guided enhanced feature O_t and count token O_{count}.

$$[O_t, O_{count}] = DeformAttn([G_t, G_{count}], \{G_r, F_r^3, F_r^2, F_r^1\}) \tag{9}$$

where $DeformAttn(a,b)$ is the multi-scale deformable attention [50], a represents content feature, b is multi-scale features.

3.3 Regression head and loss function

To obtain the density map, we use a simple regression head which consists of two 3×3 convolution layers and one 1×1 convolution layer.

$$D = RH(O_t) \tag{10}$$

where RH is the regression head.

The loss includes a loss about the density map and a loss about the learnable count token.

$$\mathcal{L} = \mathcal{L}_D(D, D^\star) + \mathcal{L}_C(O_{count}, C^\star) \tag{11}$$

where $\mathcal{L}_D$ adopts distribution matching loss proposed in[33], which supervises the density map regression and count estimation, $\mathcal{L}_C$ adopts L_1 norm ($\|\cdot\|_1$) to supervise the count token. $D^\star$ and $C^\star$ represent the ground truth of density map and count, respectively.

4 Experiments

4.1 Datasets and evaluation metrics

Dataset. The public RGBT-CC[20] dataset is adopted to evaluate our method. RGBT-CC consists of 1,030 training samples, 200 validation samples, and 800 testing ones.

Evaluation Metrics. The widely used Grid Average Mean Absolute Error (GAME)[11] and Root Mean Square Error (RMSE) are used as evaluation metrics[20, 29].

$$GAME(l) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{4^l} \mid \hat{P}_i^j - P_i^j \mid \tag{12}$$

where $\hat{P}_i^j$ represents the predicted value of the j^{th} region of the i^{th} image, P_i^j indicates the ground truth corresponding to $\hat{P}_i^j$, 4^l means the number of the divided non-overlapping regions of the image, and N is the number of paired images in testing dataset. *GAME* sums the counting errors in all the regions.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{P}_i - P_i)^2} \tag{13}$$

where $\hat{P}_i$ represents the predicted value of the i^{th} image, P_i indicates the ground truth corresponding to $\hat{P}_i$. For both *RMSE* and *GAME*, lower value means the better performance.

4.2 Implementation details

The implementation setting includes: (1) GPU (NVIDIA RTX 3090); (2) input image size (224×224); (3) train time (17 hours); (4) learning rate ($1e-5$); (5) weight decay ($1e-4$).

4.3 Comparison with state-of-the-art methods

To make quantitative comparisons, our method is compared with recent prominent approaches, including CSRNet[15], BL[22], DM-Count[33], P2PNet[26], MARUNet[23], MAN[18], CMCRL[20], TAFNet [29], MAT[41], and DEFNet[49] which are shown in Table 1. The top of the table shows six single-modal crowd counting models which are retrained by the input fusion of RGB and thermal images. The bottom of the table shows four RGB-T crowd counting models and ours. From the observation, we can conclude our method performs the best among all the methods. It achieves about 8.4%, 7.8%, 5.7%, 4.1%, 10.9% improvement over the second best result in GAME(0), GAME(1), GAME(2), GAME(3) and RMSE, respectively. The great improvement profits from the multi-modal fusion under the guidance of count token and count enhancement of a modality under the guidance of the other modality.

4.4 Ablation studies

4.4.1 Effectiveness analysis of the proposed modules

To verify the effectiveness of the proposed modules, we conduct the ablation studies. Table 2 show the result. At first, we construct a baseline model. It concatenates high-layer features of two PVT encoders and applies regression head to predict the density map and sum up.

Table 1: Comparison results of different methods on RGBT-CC benchmark dataset. **The top part:** some RGB crowd counting models are retrained by input fusion of color modality and thermal modality. **The bottom part:** some RGB-T crowd counting models. The best result is in bold.

Methods	Source	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
CSRNet[15]	CVPR2018	20.40	23.58	28.03	35.51	35.26
BL[22]	ICCV2019	18.70	22.55	26.83	34.62	32.67
DM-Count[33]	NeurIPS2020	16.54	20.73	25.23	32.23	27.22
P2PNet[26]	ICCV2021	16.24	19.42	23.48	30.27	29.94
MARUNet[23]	WACV2021	17.39	20.54	23.69	27.36	30.84
MAN[18]	CVPR2022	17.16	21.78	28.74	41.59	33.84
CMCRL[20]	CVPR2021	15.61	19.95	24.69	32.89	28.18
TAFNet[29]	ISCAS2022	12.38	16.98	21.86	30.19	22.45
MAT[41]	ICME2022	12.35	16.29	20.81	29.09	22.53
DEFNet[19]	TITS2022	11.90	16.08	20.19	27.27	21.09
Ours	BMVC2022	**10.90**	**14.81**	**19.02**	**26.14**	**18.79**

The baseline result is shown in the first line. Then, we add count-guided multi-modal fusion module and modal-guided count enhancement module based on the baseline, respectively. The result is shown in the second and the third lines, respectively. Finally, we add all the modules. The result is shown in the fourth line. By the observation, *MSTTrans* improves the performance from *GAME0* (11.62) to *GAME0* (10.91). It benefits from the better fusion which has a global common objective and multi-scale concept. *MSDTrans* improves the performance from *GAME0* (11.62) to *GAME0* (11.17). It indicates the supplementary effect of a modality on the other modality. Last, the whole model achieves a best *GAME0* (10.90), which shows the effectiveness of both modules. However, we also find that *RMSE* value in the second line achieves the best result. It suggests our future work to improve the model.

Table 2: Ablation study about modules. The best result is in bold.

Variant	Baseline	Candidate MSTTrans	MSDTrans	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
No.1	✓			11.62	16.25	20.38	27.17	19.88
No.2	✓	✓		10.91	15.26	19.88	26.99	**18.32**
No.3	✓		✓	11.22	15.20	19.42	26.30	19.75
No.4	✓	✓	✓	**10.90**	**14.81**	**19.02**	**26.14**	18.79

4.4.2 Effectiveness analysis of the count-guided multi-modal fusion design

To verify our contributions, we conduct the ablation studies about the count-guided multi-modal fusion design. There are two essential design conceptions in the module. One is the guidance of the learnable count token. The other is multi-scale strategy. Table 3 show the result. At first, we show our result in the first line. Then, we remove the learnable count token from the whole model. Finally, we replace the multi-scale token transformer with vanilla multi-head self-attention. By the observation, we find that the performance declines obviously when removing the count token. It just verifies the effectiveness of the count token. Furthermore, multi-scale concept is also effective because the performance is worse when replacing our proposed multi-scale token transformer with multi-head self-attention. Compared with both, multi-scale concept plays a more important role than the learnable

count token. It also verifies our most important contribution which introduces a token level multi-scale transformer.

Table 3: Ablation study about count guidance and multi-scale concept in count-guided multi-modal fusion module. The best result is in bold. "Ours/count" represents our model removing the learnable count token. "Ours/multi-scale" represents our model with vanilla multi-head self-attention instead of the multi-scale token transformer.

Variant	Candidate			GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
	Ours	Ours/count	Ours/multi-scale					
No.1	✓			**10.90**	**14.81**	**19.02**	**26.14**	**18.79**
No.2		✓		11.82	15.91	20.10	27.13	20.54
No.3			✓	11.82	16.39	20.89	28.37	21.73

5 Conclusions

In the paper, we propose an RGB-T multi-modal crowd counting method based on Transformer. Two-modal features are fused under the guidance of a learnable count token. Then crowd density map is predicted by a modality and guided by the other modality. To solve the large-scale variation problem, a multi-scale token transformer is proposed to diversify the receptive fields. The experimental results demonstrate a significant improvement over existing RGB-T crowd counting methods and verify the effectiveness of all the designs.

6 Acknowledgment

This work is supported by Natural Science Foundation of Anhui Province (1908085MF182) and Science Research Project for Graduate Student of Anhui Provincial Education Department (YJS20210047).

References

[1] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching Convolutional Neural Network for Crowd Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5744–5752, 2017.

[2] Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive Dilated Network with Self-Correction Supervision for Counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4594–4603, 2020.

[3] Jiwei Chen, Kewei Wang, Wen Su, and Zengfu Wang. SSR-HEF: Crowd Counting with Multi-Scale Semantic Refining and Hard Example Focusing. *IEEE Transactions on Industrial Informatics*, pages 6547–6557, 2022.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[5] Xiaoshuang Chen and Hongtao Lu. Reinforcing Local Feature Representation for Weakly-Supervised Dense Crowd Counting. *arXiv preprint arXiv:2202.10681*, 2022.

[6] Feng Dai, Hao Liu, Yike Ma, Xi Zhang, and Qiang Zhao. Dense Scale Network for Crowd Counting. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 64–72, 2021.

[7] Zizhu Fan, Hong Zhang, Zheng Zhang, Guangming Lu, Yudong Zhang, and Yaowei Wang. A Survey of Crowd Counting and Density Estimation based on Convolutional Neural Network. *Neurocomputing*, 472:224–251, 2022.

[8] Guangshuai Gao, Junyu Gao, Qingjie Liu, Qi Wang, and Yunhong Wang. CNN-Based Density Estimation and Crowd Counting: A Survey. *arXiv preprint arXiv:2003.12783*, 2020.

[9] Junyu Gao, Qi Wang, and Xuelong Li. Pcc net: Perspective crowd counting via spatial convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3486–3498, 2019.

[10] Junyu Gao, Maoguo Gong, and Xuelong Li. Congested crowd instance localization with dilated convolutional swin transformer. *Neurocomputing*, pages 94–103, 2022.

[11] Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto López-Sastre, Saturnino Maldonado-Bascón, and Daniel Onoro-Rubio. Extremely Overlapping Vehicle Counting. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 423–431. Springer, 2015.

[12] Minh Hoai and Andrew Zisserman. Talking Heads: Detecting Humans and Recognizing Their Interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 875–882, 2014.

[13] Haroon Idrees, Khurram Soomro, and Mubarak Shah. Detecting Humans in Dense Crowds Using Locally-Consistent Scale Prior and Global Occlusion Reasoning. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):1986–1998, 2015.

[14] Xiaoheng Jiang, Li Zhang, Tianzhu Zhang, Pei Lv, Bing Zhou, Yanwei Pang, Mingliang Xu, and Changsheng Xu. Density-Aware Multi-Task Learning for Crowd Counting. *IEEE Transactions on Multimedia*, 23:443–453, 2020.

[15] Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018.

[16] Dongze Lian, Xianing Chen, Jing Li, Weixin Luo, and Shenghua Gao. Locating and counting heads in crowds with a depth prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–17, 2021.

[17] Dingkang Liang, Xiwu Chen, Wei Xu, Yu Zhou, and Xiang Bai. TransCrowd: Weakly-Supervised Crowd Counting with Transformers. *Science China Information Sciences*, 65(6):1–14, 2022.

[18] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting Crowd Counting via Multifaceted Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19628–19637, 2022.

[19] Lingbo Liu, Jiajie Zhen, Guanbin Li, Geng Zhan, Zhaocheng He, Bowen Du, and Liang Lin. Dynamic Spatial-Temporal Representation Learning for Traffic Flow Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 22(11):7169–7183, 2020.

[20] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4823–4833, 2021.

[21] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5099–5108, 2019.

[22] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian Loss for Crowd Count Estimation with Point Supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6142–6151, 2019.

[23] Liangzi Rong and Chunping Li. Coarse- and Fine-Grained Attention Network with Background-Aware Loss for Crowd Density Map Estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3675–3684, 2021.

[24] Usman Sajid, Xiangyu Chen, Hasan Sajid, Taejoon Kim, and Guanghui Wang. Audio-visual transformer based crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2249–2259, 2021.

[25] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. Revisiting Perspective Information for Efficient Crowd Counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7279–7288, 2019.

[26] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking Counting and Localization in Crowds: A Purely Point-Based Framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3365–3374, 2021.

[27] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-End People Detection in Crowded Scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2325–2333, 2016.

[28] Guolei Sun, Yun Liu, Thomas Probst, Danda Pani Paudel, Nikola Popovic, and Luc Van Gool. Boosting Crowd Counting with Transformers. *arXiv preprint arXiv:2105.10926*, 2021.

[29] Haihan Tang, Yi Wang, and Lap-Pui Chau. TAFNet: A Three-Stream Adaptive Fusion Network for RGB-T Crowd Counting. *arXiv preprint arXiv:2202.08517*, 2022.

[30] Ye Tian, Xiangxiang Chu, and Hongpeng Wang. CCTrans: Simplifying and Improving Crowd Counting with Transformer. *arXiv preprint arXiv:2109.14483*, 2021.

[31] Imran Usman and Abdulaziz A Albesher. Abnormal Crowd Behavior Detection Using Heuristic Search and Motion Awareness. *International Journal of Computer Science & Network Security*, 21(4):131–139, 2021.

[32] Thirumalaisamy P Velavan and Christian G Meyer. The COVID-19 Epidemic. *Tropical Medicine & International Health*, 25(3):278, 2020.

[33] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution Matching for Crowd Counting. *Advances in Neural Information Processing Systems*, 33:1595–1607, 2020.

[34] Fusen Wang, Kai Liu, Fei Long, Nong Sang, Xiaofeng Xia, and Jun Sang. Joint CNN and Transformer Network via Weakly Supervised Learning for Efficient Crowd Counting. *arXiv preprint arXiv:2203.06388*, 2022.

[35] Fusen Wang, Jun Sang, Zhongyuan Wu, Qi Liu, and Nong Sang. Hybrid Attention Network Based on Progressive Embedding Scale-Context for Crowd Counting. *Information Sciences*, 591:306–318, 2022.

[36] Mingjie Wang, Hao Cai, Xianfeng Han, Jun Zhou, and Minglun Gong. STNet: Scale Tree Network with Multi-level Auxiliator for Crowd Counting. *IEEE Transactions on Multimedia*, pages 1–11, 2022.

[37] Mingjie Wang, Jun Zhou, Hao Cai, and Minglun Gong. CrowdMLP: Weakly-Supervised Crowd Counting via Multi-Granularity MLP. *arXiv preprint arXiv:2203.08219*, 2022.

[38] Qian Wang and Toby P Breckon. Crowd Counting via Segmentation Guided Attention Networks and Curriculum Loss. *IEEE Transactions on Intelligent Transportation Systems*, pages 15233–15243, 2022.

[39] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[40] Xing Wei, Yuanrui Kang, Jihao Yang, Yunfeng Qiu, Dahu Shi, Wenming Tan, and Yihong Gong. Scene-Adaptive Attention Network for Crowd Counting. *arXiv preprint arXiv:2112.15509*, 2021.

[41] Zhengtao Wu, Lingbo Liu, Yang Zhang, Mingzhi Mao, Liang Lin, and Guanbin Li. Multimodal Crowd Counting with Mutual Attention Transformers. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.

[42] Zhaoyi Yan, Yuchen Yuan, Wangmeng Zuo, Xiao Tan, Yezhen Wang, Shilei Wen, and Errui Ding. Perspective-Guided Convolution Networks for Crowd Counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 952–961, 2019.

[43] Shangpeng Yang, Weiyu Guo, and Yuheng Ren. CrowdFormer: An Overlap Patching Vision Transformer for Top-Down Crowd Counting. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, pages 1545–1551, 2022.

[44] Yifan Yang, Guorong Li, Dawei Du, Qingming Huang, and Nicu Sebe. Embedding Perspective Analysis into Multi-Column Convolutional Neural Network for Crowd Counting. *IEEE Transactions on Image Processing*, 30:1395–1407, 2020.

[45] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Reverse Perspective Network for Perspective-Aware Object Counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4374–4383, 2020.

[46] Xiaoyuan Yu, Yanyan Liang, Xuxin Lin, Jun Wan, Tian Wang, and Hong-Ning Dai. Frequency Feature Pyramid Network With Global-Local Consistency Loss for Crowd-and-Vehicle Counting in Congested Scenes. *IEEE Transactions on Intelligent Transportation Systems*, pages 9654–9664, 2022.

[47] Lixian Yuan, Zhilin Qiu, Lingbo Liu, Hefeng Wu, Tianshui Chen, Pei Chen, and Liang Lin. Crowd Counting via Scale-Communicative Aggregation Networks. *Neurocomputing*, 409:420–430, 2020.

[48] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 589–597. CVPR, 2016.

[49] Wujie Zhou, Yi Pan, Jingsheng Lei, Lv Ye, and Lu Yu. DEFNet: Dual-Branch Enhanced Feature Fusion Network for RGB-T Crowd Counting. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–10, 2022.

[50] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*, pages 1–12, 2020.