BMVC
Paper ID 0427 2022

RGB-T Multi-Modal Crowd Counting Based on Transformer

Zhengyi Liu*, Wei Wu, Yacheng Tan, Guanghui Zhang

Anhui University

Task

- ➢ Crowd counting can predict the distribution of crowd and estimate the number of persons in unconstraint scenes.
- ➢ RGB-Thermal (RGB-T) crowd counting has attracted a lot of attentions because thermal modality can recognize the persons in insufficient light condition.

Motivation

➢ Solve the challenge of multi-modal fusion

Existing methods fuse complementary multi-modal features with no constraint. 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.

➢ Solve the challenge of large-scale variation

Existing methods use dilated convolution to enlarge the receptive fields. We propose 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.

Method

- **count-guide multi-modal fusion**

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.

- **modal-guide count enhancement**

Counting information is further enhanced from one modality under the guidance of the other modality.

- **regression head**

It generates density map which is supervised by GT, and meanwhile count is also supervised by GT.

Experiment

➢ Ablation

From ablation study, we find the effectiveness of modules, count token and multi-scale token transformer.

Variant	Candidate Baseline	 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

Variant	Candidate Ours	 Ours/count	 Ours/multi-scale	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
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

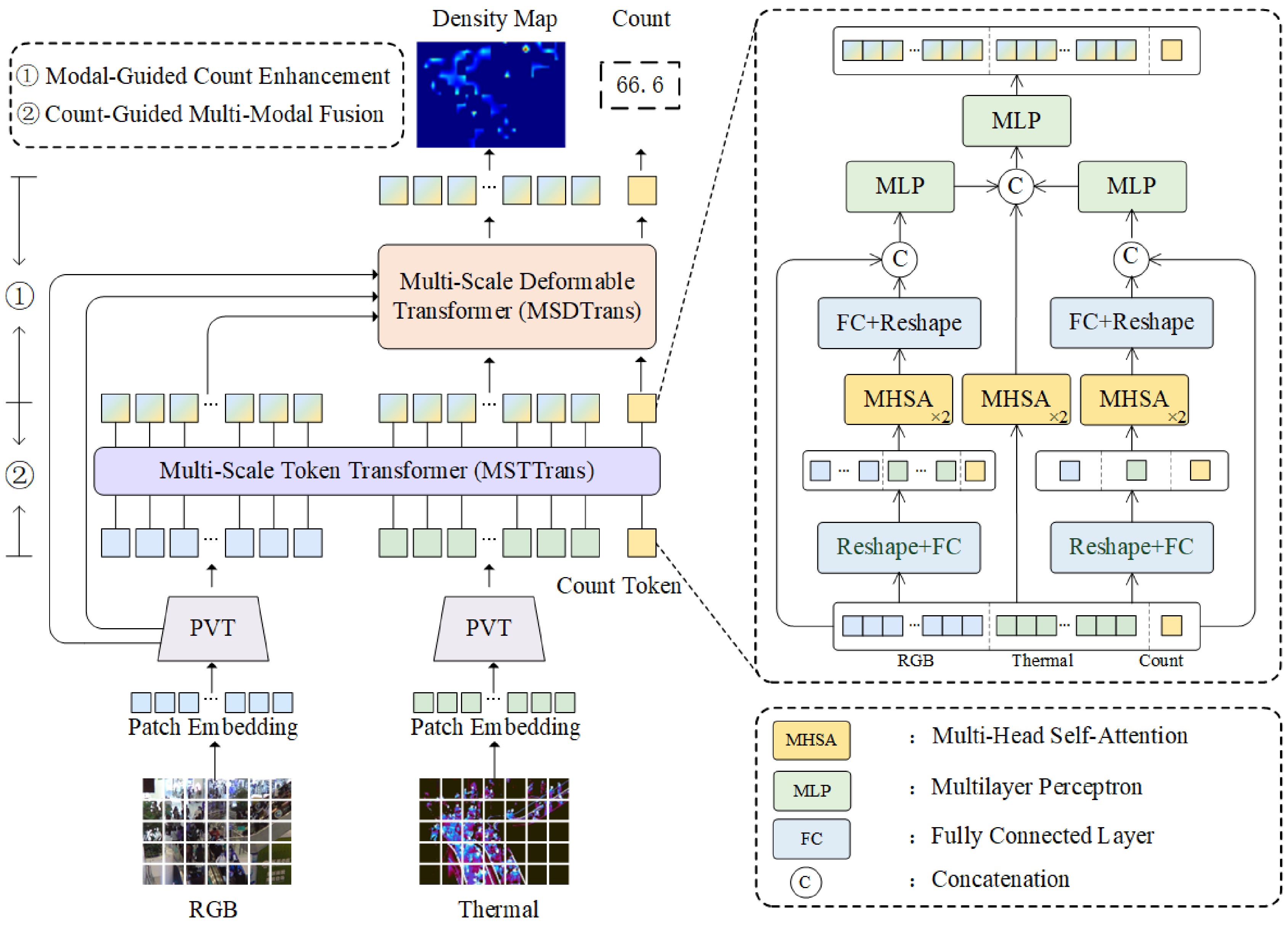
① Modal-Guided Count Enhancement
② Count-Guided Multi-Modal Fusion
Density Map
Count
66. 6
①
②
Multi-Scale Deformable Transformer (MSDTrans)
Multi-Scale Token Transformer (MSTTrans)
Count Token
PVT
PVT
Patch Embedding
Patch Embedding
RGB
Thermal
MLP
MLP
MLP
C
C
C
FC+Reshape
FC+Reshape
MHSA ×2
MHSA ×2
MHSA ×2
Reshape+FC
Reshape+FC
RGB
Thermal
Count
MHSA : Multi-Head Self-Attention
MLP : Multilayer Perceptron
FC : Fully Connected Layer
C : Concatenation

Fig 1: Main framework.

➢ Comparison

From comparison experiment, we find our method achieves the state-of-the-art performance.

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[49]	TITS2022	11.90	16.08	20.19	27.27	21.09
Ours	BMVC2022	**10.90**	**14.81**	**19.02**	**26.14**	**18.79**

Conclusion

- ➢ 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.

Code: https://github.com/liuzywen/RGBTCC

Contact: liuzywen@ahu.edu.cn