Segmentation Assisted U-shaped Transformer for Crowd Counting

Yifei Qian[1], Liangfei Zhang[1], Xiaopeng Hong[2], Carl R. Donovan[1], Ognjen Arandjelović[1]
[1]University of St. Andrews
[2] Harbin Institute of Technology

BMVC 2022

University of St Andrews

Introduction

Automated crowd counting has made remarkable progress recently in computer vision thanks to the development of CNNs. However, this application area has run into bottlenecks since CNNs, by their nature, are limited by locally attentive receptive fields and are incapable of modelling larger-scale dependencies. To address this problem, we introduce a multi-scale transformer-based crowd-counting network, termed Crowd U- Transformer (CUT) which extracts and aggregates semantic and spatial features from multiple levels. In this design, we use crowd segmentation as an attention module to gain fine-grained features. Also, we propose a loss function that better focuses on the counting performance in the foreground area. Experimental results on four widely used benchmarks are presented and our method shows state-of-the-art performances.

Proposed Method

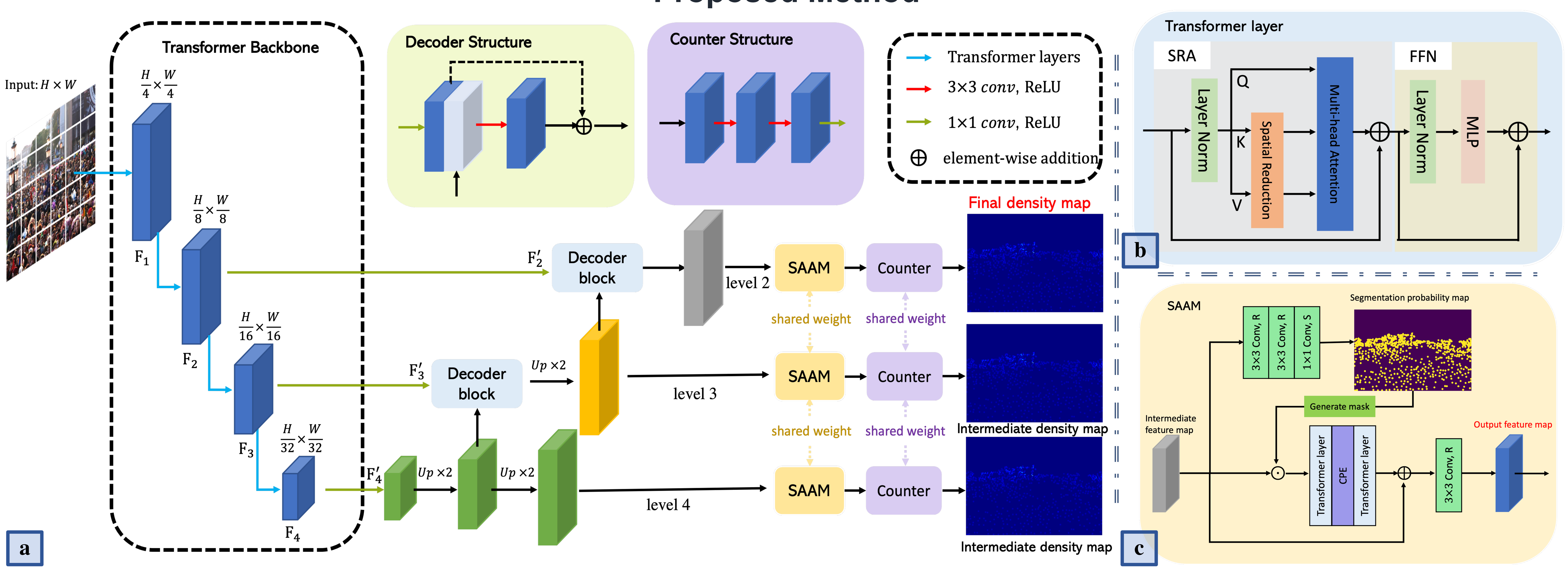

Fig.1 (a) The overview of the proposed CUT. Twins-PcPvT is adopted as the backbone network to extract multi-scale features; **(b)** The structure of the basic transformer layer in our model, proposed in PvT; **(c)** The structure of the proposed Segmentation as attention module (SAAM).

Loss Design:

- Supervisions are provided on all three levels and for each level l, we have loss function L^l,

$$L^l = L_S^l + L_R^l \quad l = 2, 3 \text{ and } 4.$$

- L_s adopts the pixel-level focal loss which is used to supervise the segmentation task,

$$L_S = -\sum_{i \in S_{gt}} l_i (1-p_i)^\gamma \log(p_i) + (1-l_i) p_i^\gamma \log(1-p_i)$$

where i denotes the pixel within S_{gt}(the ground-truth segmentation map), and $l_i$and p_i represent the actual label of the pixel and the predicted probability of that pixel being foreground, respectively; γ is the modulating factor.

- L_R is defined as follows which emphasizes the importance of counting accuracy over dense regions:

$$L_R = SL(D_p \odot S_{gt}, D_{gt} \odot S_{gt}) + \lambda \cdot L_{TV}(D_p, D_{gt})$$

where SL represents the structural loss, D_p indicates the predicted density map, $\odot$ is the element-wise multiplication, λ is the tuneable hyper-parameter, and L_{TV} is the total variation loss.

- The final loss is a combination of L_l from three levels:

$$L_{total} = \sum_{l=2}^{4} L_S^l + \alpha \sum_{l=3}^{4} L_R^l + L_R^2.$$

Key Contribution:

- We propose a 'U-shaped' design of multi-scale Transformer network for crowd counting, we refer to as Crowd U-Transformer (CUT). Our design effectively improves the model's performance in the presence of large-scale variations of objects.
- We introduce an attention module, SAAM, which leverages the crowd segmentation results and a simple transformer block to extract fine-grained features in crowd regions.
- We design a new loss function for supervising the regression which provides a significant improvement on counting accuracy over previous attempts.

Experimental Results

Model	Label level	SHA		UCF-QNRF		JHU-Crowd++		NWPU	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN (CVPR16)	density map	110.2	173.2	277	426	188.9	483.4	232.5	714.6
CSRNet (CVPR18)	density map	68.2	115.0	-	-	121.3	387.8	121.3	387.8
SANet (ECCV18)	density map	67.0	104.5	-	-	91.1	320.4	-	-
CAN (CVPR19)	density map	62.3	100.0	107.0	183.0	100.1	314.0	106.3	386.5
SFCN (CVPR19)	density map	64.8	107.5	102.0	171.4	77.5	297.6	-	-
BL (ICCV19)	point map	62.8	101.8	88.7	154.8	75.0	299.9	105.4	454.2
ASNet (CVPR20)	density map	57.8	90.1	91.6	159.7	-	-	-	-
AMRNet (ECCV20)	density map	61.5	98.3	86.6	152.2	-	-	-	-
DPN-IPSM (ACMMM20)	point map	58.1	91.7	84.7	147.2	-	-	-	-
DM-Count (NIPS20)	point map	59.7	95.7	85.6	148.3	-	-	88.4	388.6
UOT (AAAI21)	point map	58.1	95.9	84.3	142.3	60.5	252.7	87.8	387.5
S3 (IJCAI21)	point map	57.0	96.0	80.6	139.8	59.4	244.0	-	-
GL (CVPR21)	point map	61.3	95.4	84.3	147.5	59.9	259.5	79.3	346.1
D2CNet (IEEE-TIP21)	density map	57.2	93.0	81.7	137.9	73.7	292.5	85.5	361.5
CFANet (WACV21)	density map	56.1	89.6	89.0	152.3	-	-	-	-
SASNet (AAAI21)	density map	53.6	88.4	85.2	147.3	-	-	-	-
P2PNet (ICCV21)	point map	52.7	85.1	85.3	154.5	-	-	72.6	331.6
BCCT (arXiv21)	point map	53.1	82.2	83.8	143.4	54.8	**208.5**	82.0	366.9
CCTrans (arXiv21)	point map	52.3	84.9	82.8	142.3	-	-	**69.3**	**299.4**
CUT(ours)	density map	**51.9**	**79.1**	**78.4**	**135.6**	**54.3**	229.1	**69.3**	304.0

Table.1. Performance comparison with the state-of-the-art models on ShanghaiTech A, UCF- QNRF, JHU-Crowd++ and NWPU. The best and the second-best performance are shown in **bold** and underlined, respectively.

Ablation Study

Table 2. The effectiveness of multi-level supervision (MLS).

	SHA		UCF-QNRF	
	MAE	MSE	MAE	MSE
w/o MLS	54.0	88.1	79.1	138.1
Ours	**51.9**	**79.1**	**78.4**	**135.6**

Table 3. The effectiveness of SAAM comparing with the traditional 'probability as weight' (PAW) module.

	SHA		UCF-QNRF	
	MAE	MSE	MAE	MSE
w/o SAAM	54.2	85.9	82.1	144.2
PAW	53.6	82.5	80.6	146.2
Ours	**51.9**	**79.1**	**78.4**	**135.6**

Table 4. The effectiveness of proposed regression function comparing with Background Structural loss (BSL).

	SHA		UCF-QNRF	
	MAE	MSE	MAE	MSE
BSL	56.5	90.3	87.4	144.3
Ours	**51.9**	**79.1**	**78.4**	**135.6**

Acknowledgement：This work is supported by the Fundamental Research Funds for the Central Universities (AUGA5710011522). The authors would like to thank the China Scholarship Council – University of St Andrews Scholarships (No.201908060250) funds L. Zhang for her PhD.