

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 (CUT) which extracts and aggregates semantic and spatial features from multiple levels. In this design, we use crowd segmentation as an attention 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.



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} , $L^{l} = L^{l}_{S} + L^{l}_{R}$ l = 2, 3 and 4. • L_s adopts the pixel-level focal loss which is used to supervise the segmentation task,

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

where i denotes the pixel within S_{qt} (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:

$$L_{total} = \sum_{l=2}^{4} L_{S}^{l} + \alpha \sum_{l=1}^{4} L_{S}^{l}$$

Segmentation Assisted U-shaped Transformer for Crowd Counting

Yifei Qian¹, Liangfei Zhang¹, Xiaopeng Hong², Carl R. Donovan¹, Ognjen Arandjelović¹ ¹University of St. Andrews ² Harbin Institute of Technology

Introduction

Fig.1 (a) The overview of the proposed CUT. Twins-PcPvT is adopted as the backbone network to extract multi-scale features; (b) The struct

 $(-l_i)p_i^{\gamma}\log(1-p_i)$

Key Contribution:

- We propose a 'U-shape scale Transformer net counting, we refer to a Transformer (CUT). C improves the model's presence of large-scale
- We introduce an attent which leverages the cre results and a simple tra extract fine-grained fea regions.
- We design a new loss supervising the regress significant improvement accuracy over previous

 $L_R^l + L_R^2.$



	Exp	erime	ental	Res	uits				
Model	I abel level	SH	[A	UCF-0	QNRF	JHU-C	crowd++	NW	/PU
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	Μ
MCNN (CVPR16)	density map	110.2	173.2	277	426	188.9	483.4	232.5	71
CSRNet (CVPR18)	density map	68.2	115.0	-	-	121.3	387.8	121.3	38
SANet (ECCV18)	density map	67.0	104.5	-	-	91.1	320.4	-	
\leftarrow CAN (CVPR19)	density map	62.3	100.0	107.0	183.0	100.1	314.0	106.3	38
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	45
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 (ACMMN	(120) point map	58.1	91.7	84.7	147.2	-	-	-	•
$=$ $=$ $=$ \parallel \parallel DM-Count (NIPS20)	point map	59.7	95.7	85.6	148.3	-	-	88.4	38
$\overset{map}{=} \overset{u}{=} \overset{u}{=} \overset{UOT}{} (AAAI21)$	point map	58.1	95.9	84.3	142.3	60.5	252.7	87.8	38
	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	34
D2CNet (IEEE-TIP21) density map	57.2	93.0	81.7	137.9	73.7	292.5	85.5	36
CFANet (WACV21)	density map	56.1	89.6	89.0	152.3	-	-	-	
Output feature map	density map	53.6	88.4	85.2	147.3	-	-	-	• •
P2PNet (ICCV21)	point map	52.7	85.1	85.3	154.5	-	-	$\frac{72.6}{22.6}$	33
$ \begin{array}{c} \blacksquare \\ \blacksquare $	point map	53.1	<u>82.2</u>	83.8	143.4	<u>54.8</u>	208.5	82.0	36
	point map	523	84.9	82.8	42.3		-	69.3	- 29
e basic	density map density map unce comparison J-Crowd++ and <u>underlined</u> , resp = = = = = = =	51.9 with to the sectively	79.1 he state . The .	78.4 te-of-th best an	135.6 e-art n d the s	54.3 nodels second	<u>229.1</u> on Sha -best pe	69.3 anghaiT erforma	<u>30</u> . Tech nce
e basic Ablation Student	density map density map ance comparison J-Crowd++ and <u>underlined</u> , resp = = = = = = = =	51.9 with to the section of the se	79.1 he state . The .	78.4 te-of-th best an	135.6 e-art n d the s	54.3 nodels second	<u>229.1</u> on Sha -best pe	69.3 anghaiT erforma	<u>30</u> Tech nce
e basic = = = = = = = = = = = = = = = = = = =	density map density map ance comparison J-Crowd++ and <u>underlined</u> , resp = = = = = = = = = = = = = = = = = = =	51.9 with to the section of the se	79.1 he state 7.	78.4 te-of-th best an	$\frac{112.5}{135.6}$ $e-art n$ $d the s$ $= = = =$	54.3 nodels second	<u>229.1</u> on Sha -best pe	69.3 anghaiT erforma 	30 Tech nce
e basic n of multi-	density map density map ance comparison J-Crowd++ and <u>underlined</u> , resp = = = = = = = = = = = = = = = = = = =	51.9 with to the section of the se	79.1 he state 7.	78.4 te-of-th best an	$\frac{112.5}{135.6}$ $e-art n$ $d the s$ $= = = =$	54.3 nodels second SE	229.1 on Sha -best pe = = = = UCF-0 MAE	69.3 anghaiT erforma 	<u>30</u> Tech nce
e basic of multi- rowd CUT(ours) Table.1. Performat UCF- QNRF, JHU shown in bold and ===== Ablation Stu of multi-level s	density map density map ance comparison J-Crowd++ and <u>underlined</u> , resp = = = = = = = = = = = = = = = = = = =	51.9 with t NWPU ectively	79.1 he state . The	78.4 te-of-th best an = = = = = = = = = = = = = = = = = = =	$\frac{112.5}{135.6}$ e-art n d the s = = = = = = = = = = = = = = = = = = =	54.3 nodels second SE	<u>229.1</u> on Sha best pe <u>UCF-0</u> MAE 79.1	69.3 anghaiT erforma QNRF MSE 138.	30 Tech nce
e basic of multi- trowd U- UCF- QNRF, JHU Shown in bold and E = = = = = = = = = = = = = = = = = = =	density map ance comparison J-Crowd++ and <u>underlined</u> , resp ■ ■ ■ ■ ■ ■ ■ idy fectiveness upervision	51.9 with t NWPU ectively	79.1 he sta . The	78.4 te-of-th best an = = = = = = = = = = = = = = = = = = =	135.6 e-art n d the s SHA E MS 88 70	54.3 nodels second SE	229.1 on Sha best pe UCF-0 MAE 79.1	69.3 anghaiT erforma QNRF MSE 138.1	$\frac{30}{1}$
e basic Table.1. Performa UCF- QNRF, JHU shown in bold and = = = = = Ablation Stu Table 2. The effectively	density map density map ance comparison J-Crowd++ and <u>underlined</u> , resp ■ and y fectiveness upervision ►	51.9 with t NWPU ectively = = = =	79.1 he sta . The /ILS	78.4 te-of-th best an = = = = = S MAE 54.0 51.9	135.6 e-art n d the s SHA E MS 88 88 79	54.3 nodels second SE .1 .1 .1	229.1 on Sha best pe UCF-0 MAE 79.1 78.4	69.3 anghaiT erforma QNRF MSE 138.1 135.0	<u>30</u> Tech nce E 1 6
e basic of multi- crowd U- e ffectively basic D- cut(ours) Table.1. Performation UCF- QNRF, JHU shown in bold and E = = = = Ablation Stu (MLS).	density map density map ance comparison J-Crowd++ and <u>underlined</u> , resp = = = = = = = = = = = = = = = = = = =	51.9 with t NWPU ectively = = = =	79.1 he sta . The /ILS	78.4 te-of-th best an	135.6 e-art n d the s SHA E MS 88 88 79	54.3 nodels second SE .1 .1 .1	229.1 on Sha best pe UCF-0 MAE 79.1 78.4	69.3 anghaiT erforma QNRF MSE 138.1 135.0	<u>30</u> Tech nce E 1 6
Image: Contrains (arXiv21) CUT(ours) Table.1. Performation UCF- QNRF, JHU shown in bold and Image: Shown in bold and	density map density map ance comparison J-Crowd++ and <u>underlined</u> , resp = = = = = = = = = = = = = = = = = = =	51.9 with t NWPU ectively = = = = w/o N Ours UCF	79.1 he sta . The	78.4 te-of-th best an	135.6 e-art n d the s SHA E MS 88 88 79	54.3 nodels second SE .1 .1 .1	229.1 on Sha best pe UCF-0 MAE 79.1 78.4	69.3 anghaiT erforma QNRF MSE 138.1 135.0	<u>30</u> Tech nce
CC Trans (arXiv21) CUT(ours) Table.1. Performation UCF- QNRF, JHU shown in bold and shown in	density map density map ance comparison J-Crowd++ and <u>underlined</u> , resp = = = = = = = = = = = = = = = = = = =	51.9 with t NWPU ectively w/o N Ours UCF MAF	79.1 he sta . The	78.4 te-of-th best an = = = = = S MAE 54.0 51.9 F SF	135.6 e-art n d the s SHA E MS 88 79 79	54.3 nodels second SE 3.1 3.1 5.1 5.1	229.1 on Sha best pe UCF-0 MAE 79.1 78.4	69.3 anghaiT erforma QNRF MSE 138.1 135.0	30 Tech nce E 1 6
CC Trans (arXiv21) CUT(ours) Table.1. Performation UCF- QNRF, JHU shown in bold and shown in	density map density map ance comparison J-Crowd++ and <u>underlined</u> , resp = = = = = = = idy fectiveness upervision ► - SHA MAE MSE 54.2 95.0	51.9 with t NWPU ectively w/o N Ours UCF MAE	79.1 he states . The	78.4 te-of-th best an	$\frac{112.5}{135.6}$ $\frac{1}{135.6}$ $\frac{1}{135.6$	54.3 nodels second SE 3.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5	229.1 on Sha best per UCF-0 MAE 79.1 78.4	69.3 anghaiT erforma QNRF MSE 138. 135.0 ctivener ctivener	$\frac{30}{1}$
CC Trans (arXiv21) CUT(ours) Table.1. Performation UCF- QNRF, JHU shown in bold and shown in bo	P = P	51.9 with t NWPU ectively w/o N Ours UCF MAE 82.1	79.1 he states . The	78.4 te-of-th best an	I = 2.5 $I = 3$ I	54.3 nodels second SE 3.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5	229.1 on Sha best per UCF-C MAE 79.1 78.4	69.3 anghaiT anghaiT arforma QNRF MSE 138.1 135.0 ctivene ctivene cing wit	$\frac{30}{1}$
e basic a of multi- crowd U- e ffectively nce in the ns of objects. le, SAAM, mentation block to cUT(ours) Table.1. Performa UCF- QNRF, JHU shown in bold and a Ablation Stu Table 2. The eff of multi-level s (MLS). w/o SAAM PAW	density map density map ance comparison J-Crowd++ and <u>underlined</u> , resp ■ ■ ■ ■ ■ ■ ■ Idy fectiveness upervision ► SHA MAE MSE 54.2 85.9 53.6 82.5	51.9 with t NWPU ectively ectively w/o N Ours UCF MAE 82.1 80.6	79.1 he states . The	78.4 te-of-th best an 54.0 54.0 51.9 F SE 52 5.2	135.6 135.6 e-art n d the s d the s SHA SHA SHA 88 79 Tal of S the of S the	54.3 nodels second second SE 1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5	$\frac{229.1}{01}$ on Sha best per UCF-C MAE 79.1 78.4	69.3 anghaiT erforma QNRF MSE 138. 135. ctivene cing wit obabilit	30 Fech nce E 1 6 ss th ty
e basic a of multi- trowd U- a effectively mentation block to block to cut(ours) Table.1. Performa UCF- QNRF, JHU shown in bold and a a a blation Stu Table 2. The effectively mentation block to cut(ours) Table.1. Performa UCF- QNRF, JHU shown in bold and a a block to cut(ours) Table 2. The effectively mentation cut(ours) Table 2. The effectively mentation cut(ours) mentation cut(ours) Table 2. The effectively mentation cut(ours) mentation	density map density map ance comparison J-Crowd++ and underlined, resp = = = = = = = = = = = = = = = = = = =	51.9 with t NWPU ectively ectively w/o N Ours UCF MAE 82.1 80.6 78.4	79.1 he sta . The	78.4 te-of-th best an	135.6 135.6 e-art d <	54.3 nodels second second SE 1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5	229.1 on Sha best pe ucr-(MAE 79.1 78.4	69.3 anghaiT erforma QNRF MSE 138. 135. ctivene cing wit obabilit module	30 Tech nce E 1 6 ss th ty e.
e basic basic a of multi- browd U- cut (ours) Table.1. Performation UCF- QNRF, JHU shown in bold and a addition Stute Ablation Stute Table 2. The effort of multi-level stopping (MLS). w/o SAAM PAW Ours	IncecomparisonJ-Crowd++andunderlined, resp $=$ $=$ udyfectivenessupervisionSHAMAEMSE54.285.953.682.551.979.1	51.9 with t NWPU ectively ectively w/o N Ours UCF MAE 82.1 80.6 78.4	79.1 he sta . The	78.4 te-of-th best an S MAE 54.0 51.9 F SE 5.2 5.6	135.6 e-art m d the s d the s SHA SHA 88 79 Tal<	54.3 nodels second second SE .1 .1 .1 .1 .1 .1 .1 .1 .1 .1 .1 .1 .1	229.1 on Sha best pe ucr-(MAE 79.1 78.4 The effector onal 'pro- (PAW)	69.3 anghaiT erforma QNRF MSE 138. 135. ctivene cing wit obabilit module	30 Fech nce
e basic a of multi- trowd U- a effectively nce in the ns of objects. lle, SAAM, mentation c block to c UT(ours) Table.1. Performa UCF- QNRF, JHU shown in bold and ====== Ablation Stu Table 2. The effort of multi-level s (MLS). w/o SAAM PAW Ours	Image: Point Ling density mapince comparison U-Crowd++ and underlined, respInderlined, respIndyindyifectiveness upervisionifectiveness upervisionSHAMAEMAESHA54.253.682.551.979.1	51.9 with t NWPU ectively ectively w/o N Ours UCF MAE 82.1 80.6 78.4	79.1 he state . The	78.4 te-of-th best an S MAE 54.0 51.9 F SE 5.2 5.6 S S S S S S S S S S S S S S S S S S S	135.6 e-art n d the s SHA E MS E MS 88 79 79 79	54.3 nodels second second SE 1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5	$\frac{229.1}{01}$ on Sha -best per UCF-(MAE 79.1 78.4 The effect onal 'pre (PAW)	69.3 anghaiT anghaiT arforma QNRF MSE 138. 138. 135. ctivene cing wit obabilit modul	30 Fech nce
e basic a of multi- browd U- a effectively nce in the ns of objects. lle, SAAM, nentation crowd cor cor cut(ours) Table.1. Performa UCF- QNRF, JHU shown in bold and a Ablation Stu Table 2. The eff of multi-level s (MLS). w/o SAAM PAW Ours Table 4. The eff	Imagedensity mapance comparisonJ-Crowd++ andunderlined, resp $=$ = = = = = = = = = = = = = = = = = =	51.9 with t NWPU ectively ectively w/o N Ours UCF MAE 82.1 80.6 78.4	79.1 he sta . The	78.4 te-of-th best an MAE 54.0 54.0 51.9 F 52 5.6 5.6	135.6 e-art n d the s SHA E MS E MS 88 79 79 79	54.3 nodels second Second SE 3.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5	229.1 on Sha best pe UCF-(MAE 79.1 78.4 The effector onal 'pro (PAW)	69.3 anghaiT anghaiT arforma QNRF MSE 138. 138. 135. Ctivene cing wit obabilit module	30 Fech nce
e basic a of multi- browd U- a effectively nce in the ns of objects. de, SAAM, mentation cor wd currans (arXiV21) CUT(ours) Table.1. Performat UCF- QNRF, JHU shown in bold and Ablation Stu Table 2. The eff of multi-level st (MLS). w/o SAAM PAW Ours Table 4. The eff proposed regres	Image: Point map density map ance comparison J-Crowd++ and underlined, resp = = = = = = = = = = = = = = = = = = =	51.9 with t NWPU ectively ectively w/o N Ours UCF MAE 82.1 80.6 78.4	79.1 he sta . The	78.4 te-of-th best an MAE 54.0 51.9 F 52 5.6 5.6 SE 5.6	135.6 e-art m d the s SHA SHA SHA MS SHA MS SHA Image: Shape of state of	54.3 nodels second SE 3.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5	229.1 on Sha best pe UCF-0 MAE 79.1 79.1 78.4	69.3 anghaiT arforma QNRF MSE 138. 135. ctivene cing wit obabilit module MSE	$\frac{30}{2}$ Fech ince $\frac{2}{1}$ $\frac{1}{6}$ $\frac{1}{5}$ $\frac{1}{5}$
e basic a basic basic cutrians (arXiv21) CUT(ours) Table.1. Performat UCF- QNRF, JHU shown in bold and a Ablation Stu Ablation Stu Table 2. The eff of multi-level s (MLS). w/o SAAM PAW Ours Correction block to crowd Sor h provides a proposed regres comparing with	Image: Point map density map ance comparison J-Crowd++ and underlined, resp = = = = = = = = = = = = = = = = = = =	51.9 with t NWPU ectively ectively w/o N Ours UCF MAE 82.1 80.6 78.4	79.1 he sta . The . The . . . The . . . <td>78.4 te-of-th best an S MAE 54.0 51.9 F 52 5.6 5.6 5.6</td> <td>135.6 e-art m d the s d the s SHA SHA SHA MS A MS A MS MS MS 90</td> <td>54.3 nodels second SE 3.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5</td> <td>229.1 on Sha best per UCF-C MAE 79.1 78.4 The effector onal 'pro (PAW) UCF-C MAE 87.4</td> <td>69.3 anghaiT arforma </td> <td>$\frac{30}{2}$</td>	78.4 te-of-th best an S MAE 54.0 51.9 F 52 5.6 5.6 5.6	135.6 e-art m d the s d the s SHA SHA SHA MS A MS A MS MS MS 90	54.3 nodels second SE 3.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5	229.1 on Sha best per UCF-C MAE 79.1 78.4 The effector onal 'pro (PAW) UCF-C MAE 87.4	69.3 anghaiT arforma 	$\frac{30}{2}$
e basic a of multi- crowd U- a effectively ne ein the ns of objects. de, SAAM, e block to cor wd Table 1. Performat UCF- QNRF, JHU shown in bold and Ablation Stu Table 2. The eff of multi-level s (MLS). W/o SAAM PAW Ours Table 4. The eff proposed regres comparing with Structured lose (Joint Imp density map ince comparison J-Crowd++ and underlined, resp = = = = = = = = = = = = = = = = = = =	51.9 with t NWPU ectively ectively w/o N Ours UCF MAE 82.1 80.6 78.4	79.1 he sta . The . . .	78.4 te-of-th best an S MAE 54.0 51.9 F 56.5 56.5	135.6 e-art n d the s d the s SHA SHA SHA MS A MS A SHA MS A A A A A A A A A A A B A B B B B B B B B B B B B B B B B B	54.3 nodels second second SE 3.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5	229.1 on Sha best per UCF-(MAE 79.1 78.4 The effector onal 'pro (PAW) UCF-(MAE 87.4	69.3 anghaiT arforma anghait a	$\frac{30}{2}$ Fech nce $\frac{2}{1}$ $\frac{1}{6}$ $\frac{1}{5}$ $$