ACM multimedia
Seattle,USA Oct 12-16 2020

Towards Unsupervised Crowd Counting via Regression-Detection Bi-knowledge Transfer

Yuting Liu[1], Zheng Wang[2], Miaojing Shi[3], Shin'ichi Satoh[2], Qijun Zhao[1], Hongyu Yang[1]
[1]College of Computer Science, Sichuan University,
[2]National Institute of Informatics, [3]King's College London

ACM multimedia
Seattle,USA Oct 12-16 2020

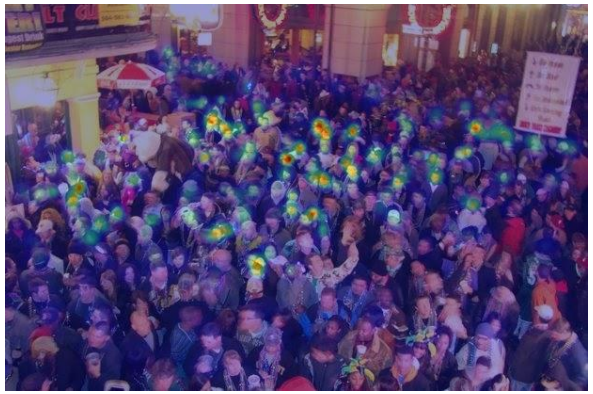
Crowd Counting(CC)

- Estimate the number of people in a crowd scene

Two typical methods

How many persons ???

By density regression



Density Map

By head detection

Head Detections

ACM multimedia
Seattle,USA Oct 12-16 2020

Bottom Necks to Open-world CC

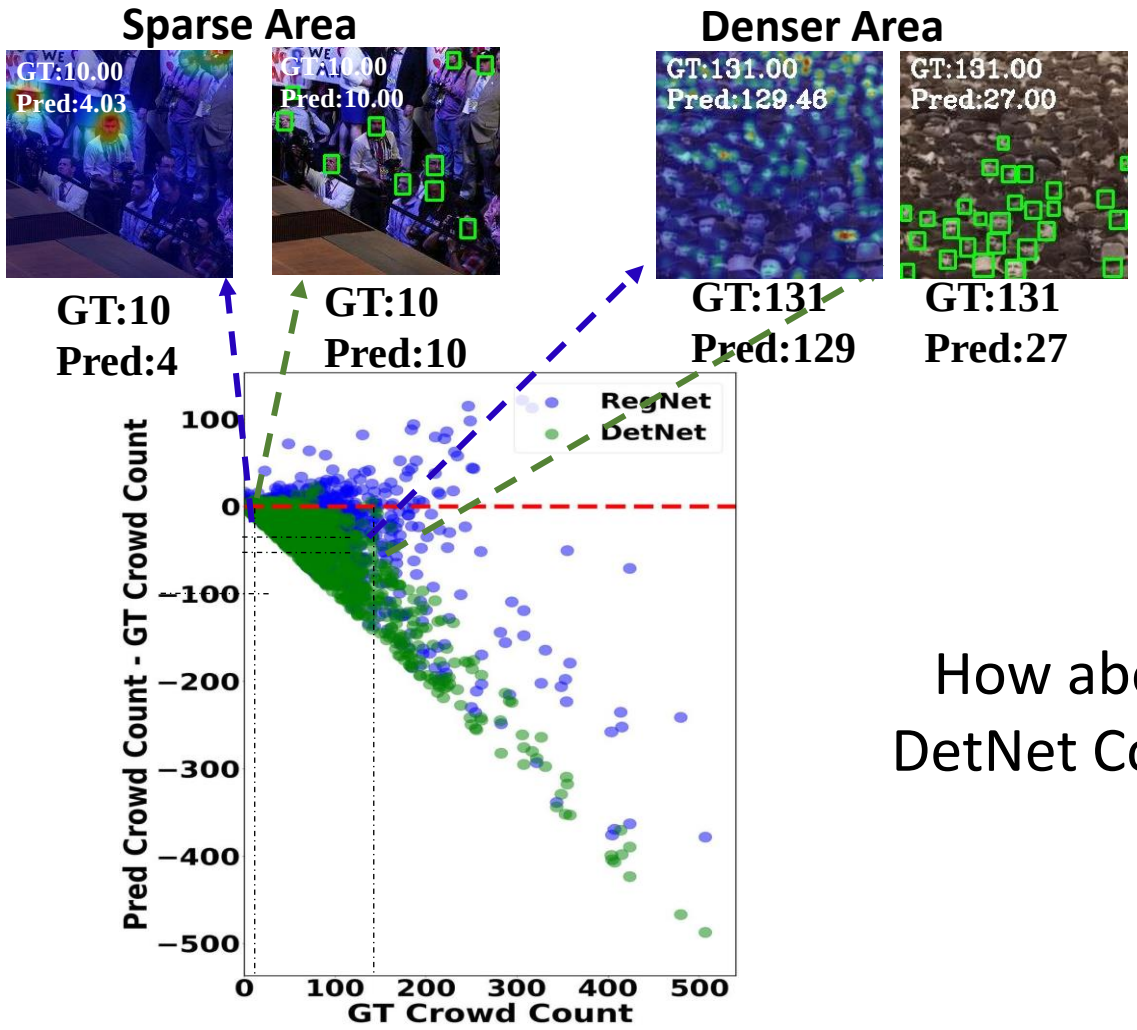
- Huge labelling Cost
- Scene-dependent Crowd Counter

Exhausted Dot Annotation

Scene variance

- What do we care: Alleviating the annotation burden
- In this work: Unsupervised cross-scene transferring

Observation : complementary property

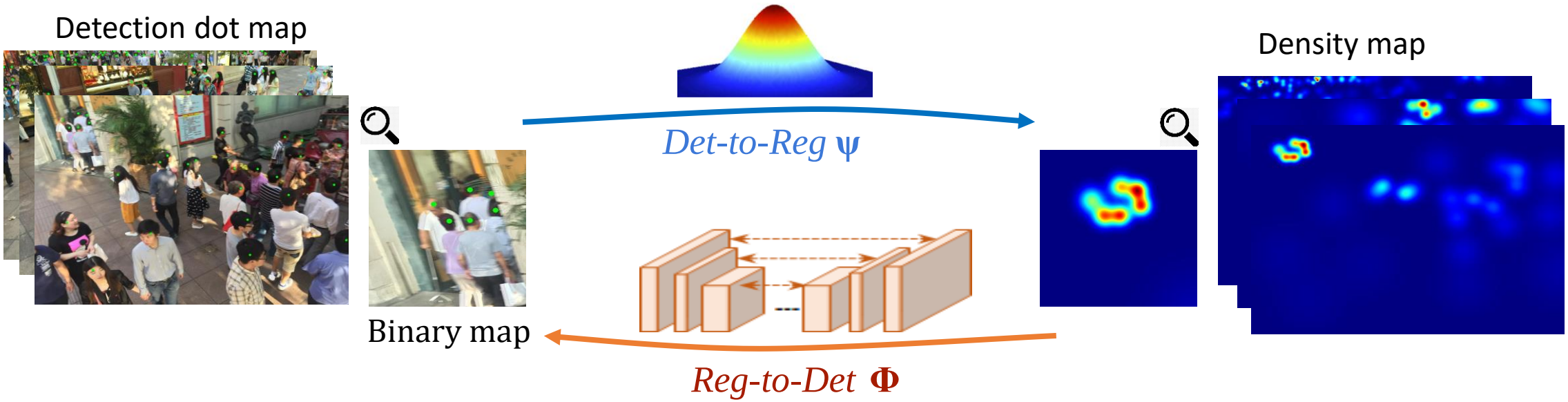


How about letting RegNet and DetNet Co-teaching each other in the Target?

ACM multimedia
Seattle,USA Oct 12-16 2020

Motivation: Reg-Det Bi-knowledge Transfer

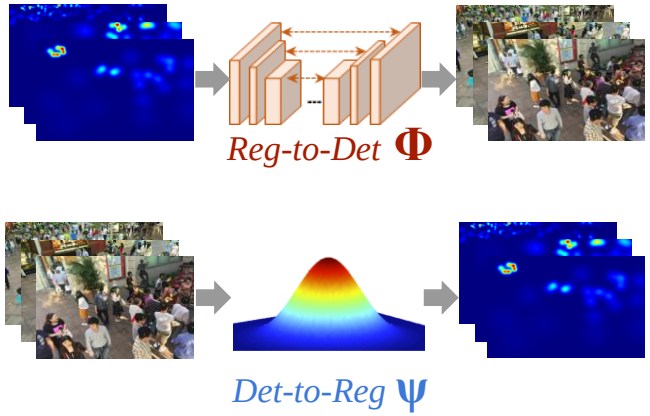
- How to co-teach RegNet & DetNet?
  - **--Scene agnostic <u>mutual transformations modelling</u> in Source**
  - **-- Reg-Det bi-knowledge transfer in Target**



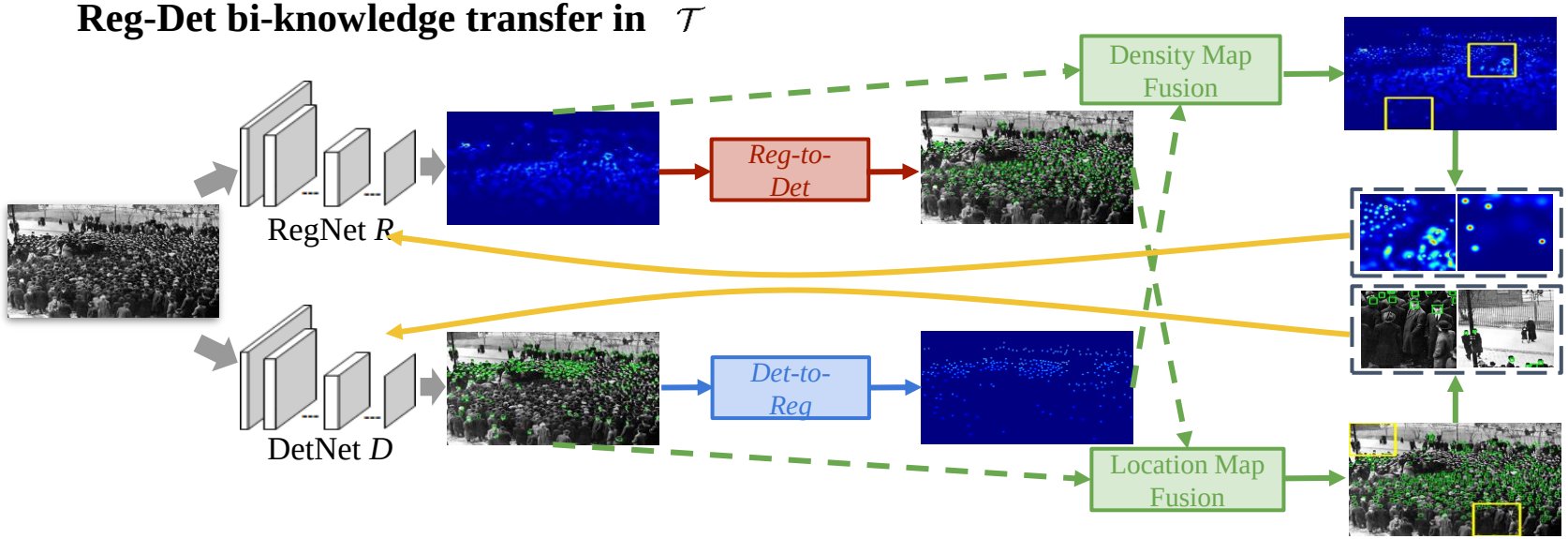
ACM multimedia
Seattle,USA Oct 12-16 2020

Our Approach: overview

Reg-Det mutual transformations modelling in $\mathcal{S}$



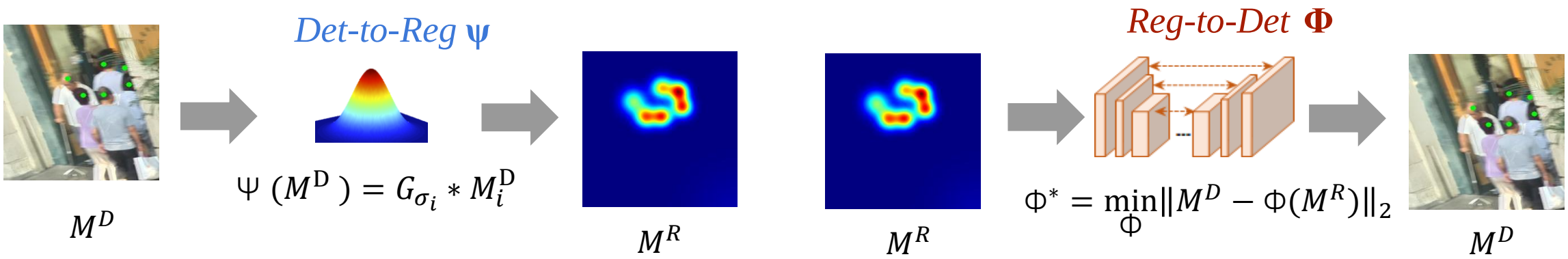
Reg-Det bi-knowledge transfer in $\mathcal{T}$



Step 1: Density and location inference
Step 2: Pseudo ground truth generation
Step 3: Self-supervised fine-tuning

ACM multimedia
Seattle,USA Oct 12-16 2020

Our Approach: **mutual transformations modelling**



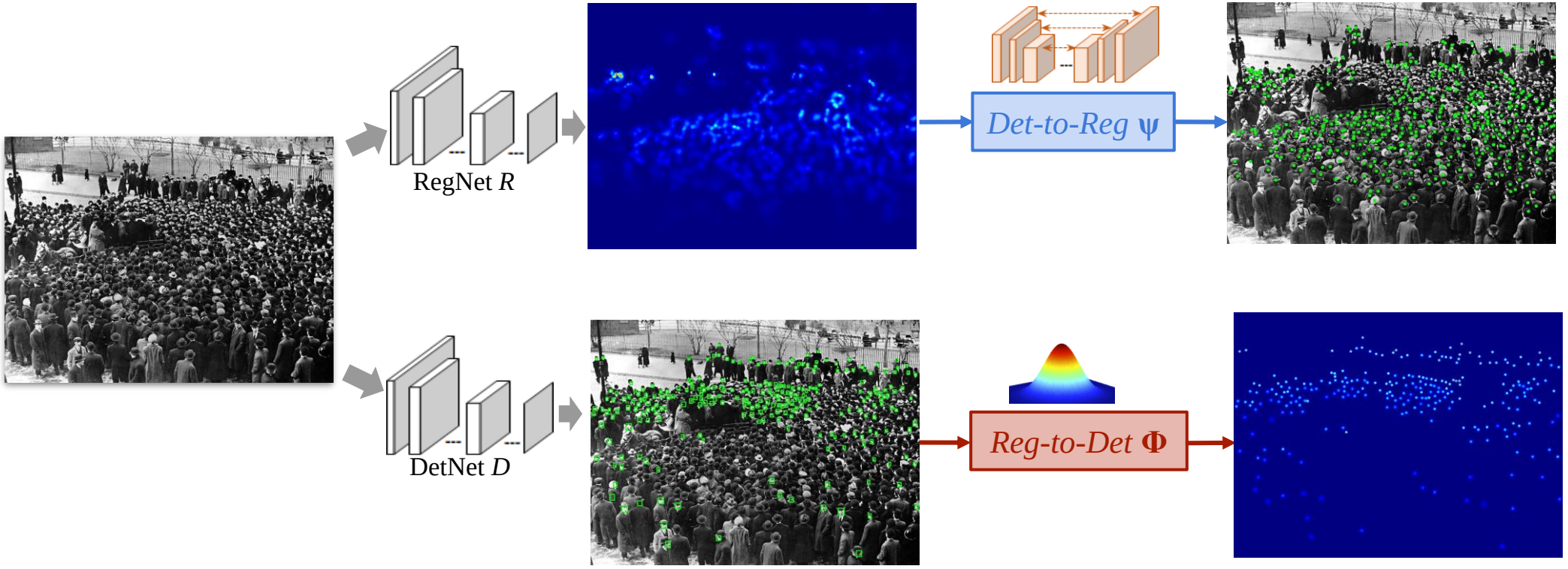
$$loss = l_{focal-mse} + l_{dms-ssim}$$

$$l_{focal-mse} = \kappa_i (1 - p_i)^{\gamma} (M^D - \Phi(M^R)^2$$

$$p_i = \begin{cases} \frac{1}{1+e^{-(\Phi(M_i^R))}}, & M_i^D = 1 \\ 1 - \frac{1}{1+e^{-(\Phi(M_i^R))}}, & o.w. \end{cases} \quad , \quad \kappa_i = \begin{cases} 1, & M_i^D = 1 \\ 0.1, & o.w. \end{cases}$$

ACM multimedia
Seattle,USA Oct 12-16 2020

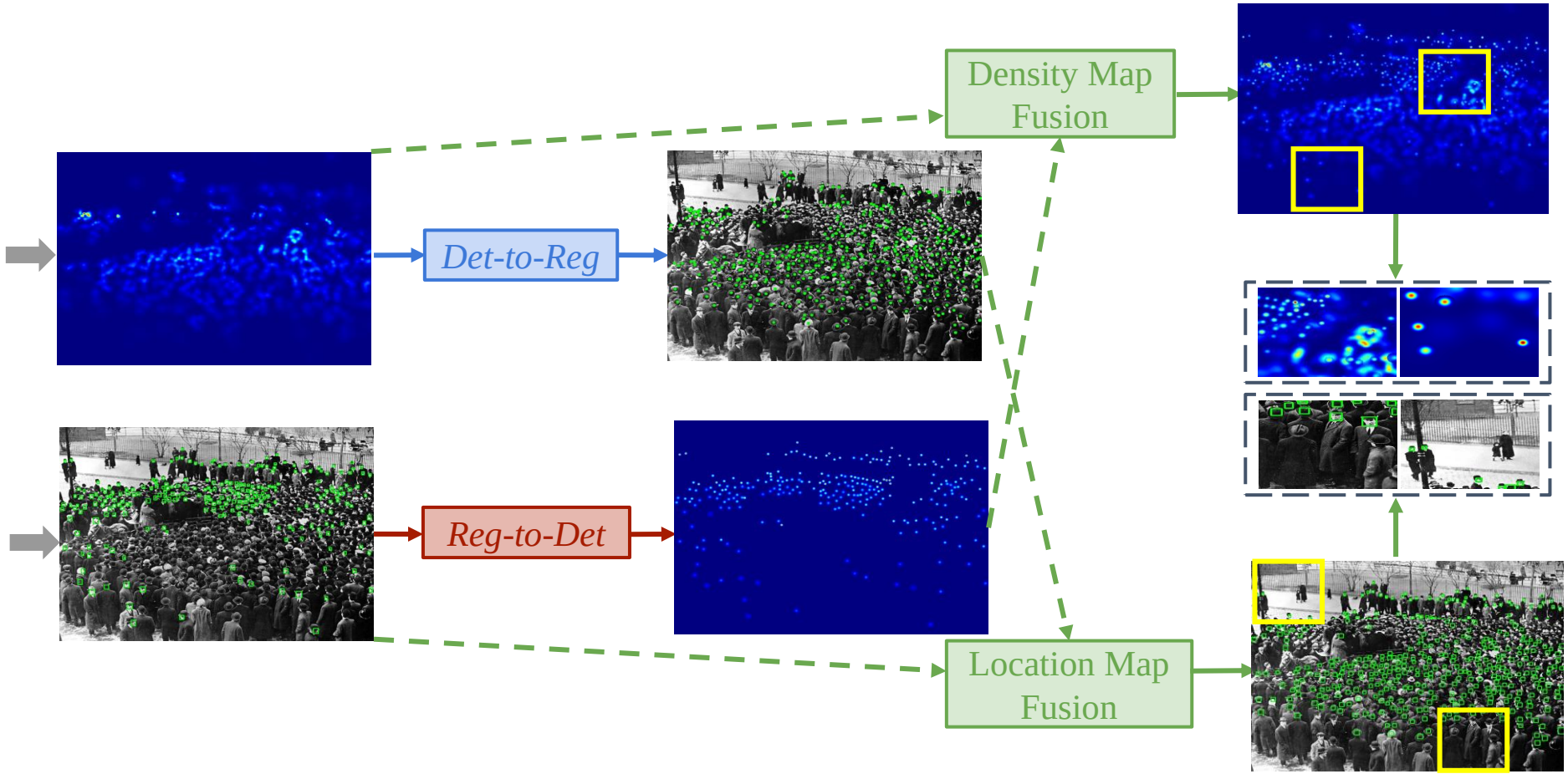
Our Approach: **Reg-Det bi-knowledge transfer**



Step 1: Density and location inference

ACM multimedia
Seattle,USA Oct 12-16 2020

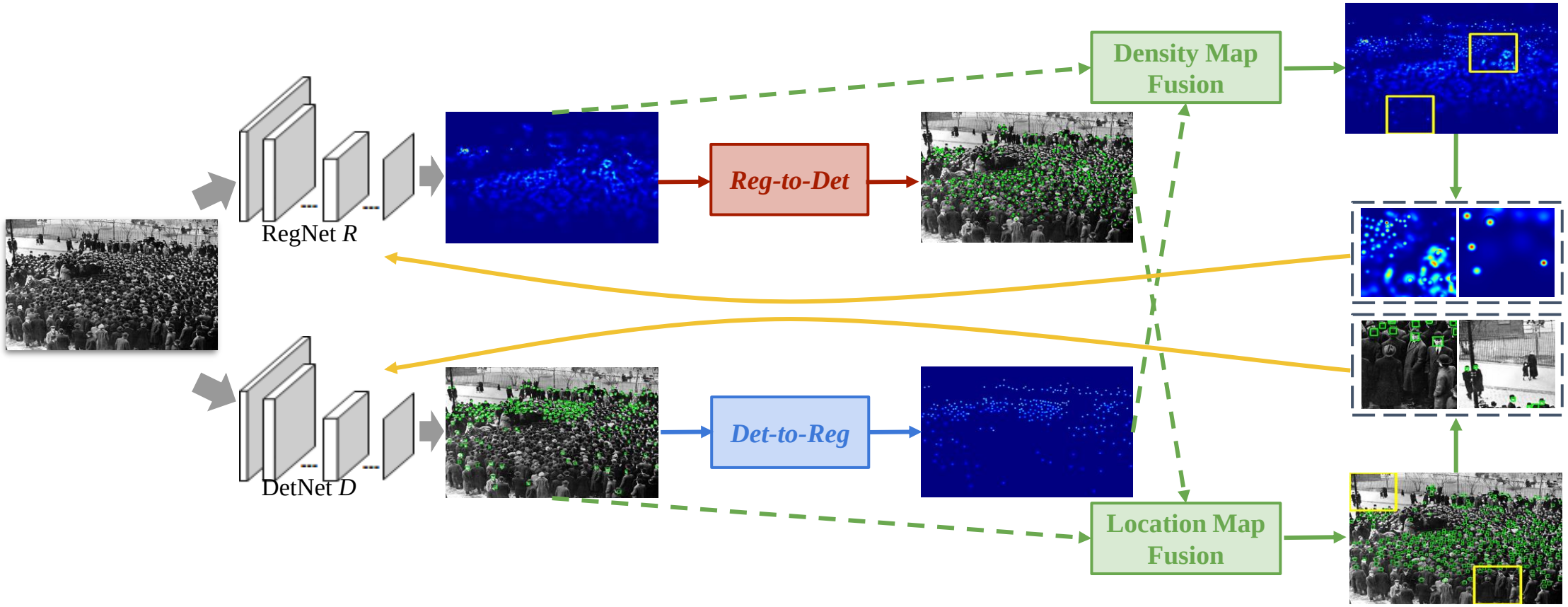
Our Approach: **Reg-Det bi-knowledge transfer**



Step 2: Pseudo ground truth generation

ACM multimedia
Seattle,USA Oct 12-16 2020

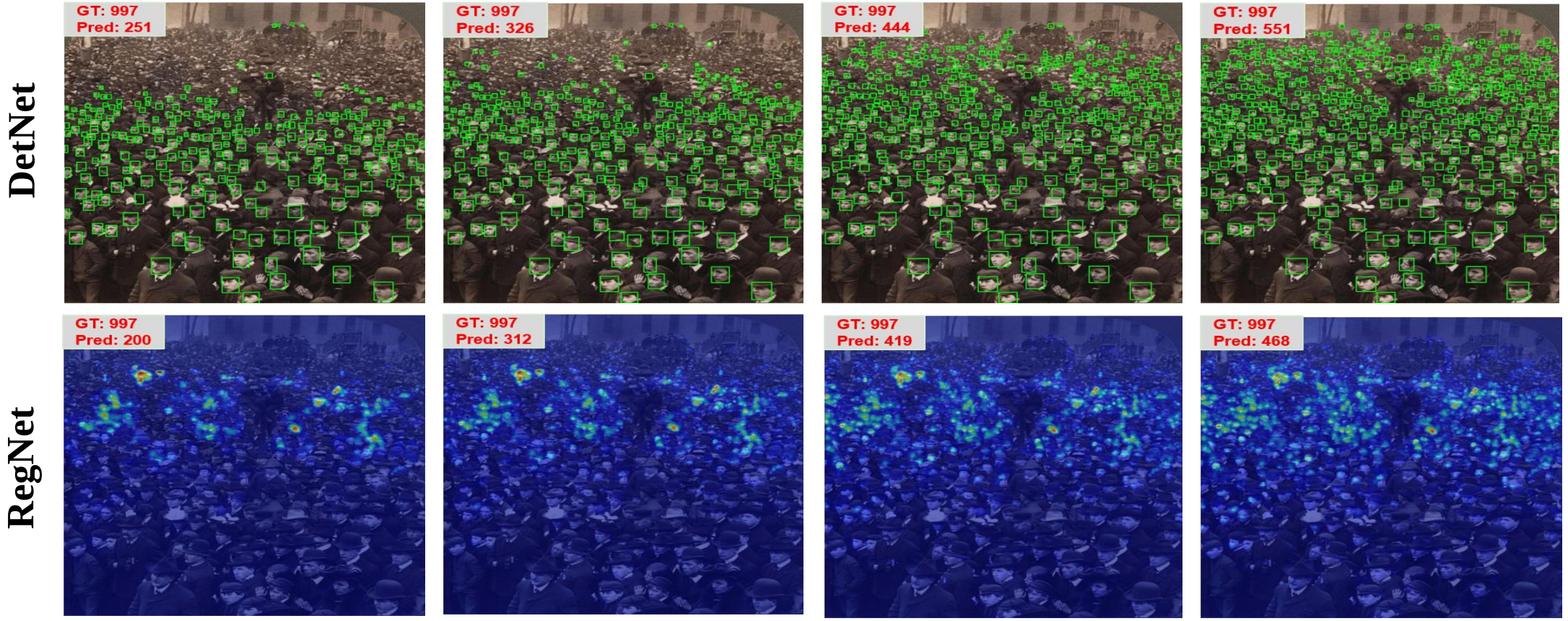
Our Approach: **Reg-Det bi-knowledge transfer**



Step 3: Self-supervised fine-tuning

ACM multimedia
Seattle,USA Oct 12-16 2020

Experiments



ACM multimedia
Seattle,USA Oct 12-16 2020

Experiments

Table 1: **Comparisons with the state-of-the-art methods** in the transfer setting. Syn denotes transfer learning using a large size synthetic dataset. Real denotes transfer learning using a real source dataset. The methods with $*$ indicate the source is the synthetic dataset GCC.

Method		A→B			A→C			A→Q			B→A			B→Q		
		MAE↓	MSE↓	mAP↑	MAE↓	MSE↓	mAP↑	MAE↓	MSE↓	mAP↑	MAE↓	MSE↓	mAP↑	MAE↓	MSE↓	mAP↑
Cycle GAN* [44]	Syn	25.4	39.7	–	404.6	548.2	–	257.3	400.6	–	143.3	204.3	–	257.3	400.6	–
SE Cycle GAN* [35]	Syn	19.9	28.3	–	373.4	528.8	–	230.4	384.5	–	123.4	193.4	–	230.4	384.5	–
SE+FD* [5]	Syn	16.9	24.7	–	–	–	–	221.2	390.2	–	129.3	187.6	–	221.2	390.2	–
MCNN [41]	Real	85.2	142.3	–	397.7	624.1	–	–	–	–	221.4	357.8	–	–	–	–
D-ConvNet-v1 [28]	Real	49.1	99.2	–	364	545.8	–	–	–	–	140.4	226.1	–	–	–	–
L2R [21]	Real	–	–	–	337.6	434.3	–	–	–	–	–	–	–	–	–	–
SPN+L2SM [37]	Real	21.2	38.7	–	332.4	425.0	–	227.2	405.2	–	126.8	203.9	–	–	–	–
RegNet [18]	Real	21.65	37.56	–	419.53	588.90	–	198.71	329.40	–	148.94	273.86	–	267.26	477.61	–
DetNet [20]	Real	55.49	90.03	0.571	703.72	941.43	0.258	411.72	731.37	0.404	242.76	400.89	0.489	411.72	731.37	0.404
Ours	Real	**13.38**	29.25	**0.757**	368.01	518.92	**0.518**	**175.02**	**294.76**	**0.546**	**112.24**	218.18	**0.661**	**211.30**	**381.92**	**0.535**

ACM multimedia
Seattle,USA Oct 12-16 2020

Experiments

➢How to learn a good Reg-to-Det Φ?

Table 2: Investigation on training the Reg-to-Det transformer

Solution	Loss	A→Q	A→B
Encoder-decoder	L_{MSE}	0.347	0.482
	$L_{Focal-MSE}$	0.423	0.599
	$L_{Focal-CE}$	0.399	0.554
	L_{Φ}	0.448	0.613

➢How to generate samples for fine-tuning?

Table 3: Ablation study on generating pseudo GT

B→A	Regression		Detection		
	MAE↓	MSE↓	MAE↓	MSE↓	mAP↑
Ours *w/o Fusion*	146.78	275.35	251.75	407.04	0.467
Ours *w/ RS*	152.19	281.05	165.80	302.90	0.609
Ours	112.24	218.18	124.12	222.42	0.661

➢How to adapt the model to the target?

Table 4: Ablation study on self-supervised learning iterations

B→A	Regression		Detection		
	MAE↓	MSE↓	MAE↓	MSE↓	mAP↑
Iteration 0	148.94	273.86	242.76	400.89	0.489
Iteration 1	136.81	258.38	172.36	305.04	0.598
Iteration 2	125.86	228.31	138.18	251.75	0.621
Iteration 3	114.06	222.55	126.86	231.45	0.658
Iteration 4	112.24	218.18	124.12	222.42	0.661
Iteration 5	113.06	220.17	125.12	213.26	0.665

ACM multimedia
Seattle,USA Oct 12-16 2020

Conclusion

- New unsupervised crowd counting scheme via regression-detection bi-knowledge
- Formulate the mutual transformations between the output of regression and detection models to enable knowledge distillation between them
- Co-training regression and detection models using pseudo labels in an iterative self-supervised way

Thank you!

732