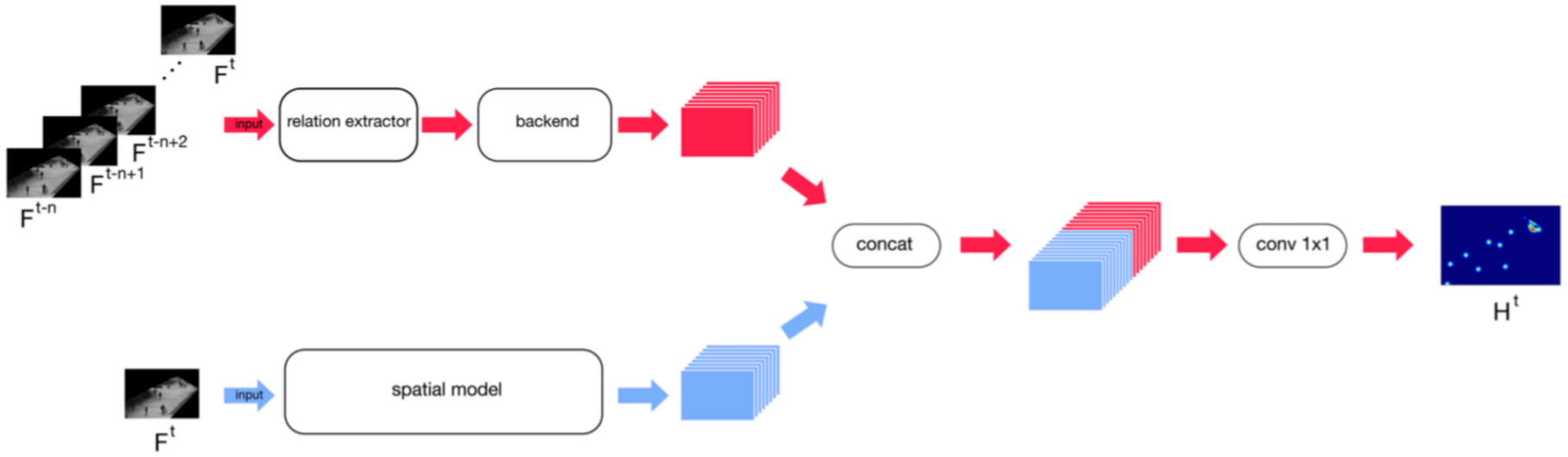
PHNet: Parasite-Host Network for Video Crowd Counting

*Shiqiao Meng, *Jiajie Li✉, Weiwei Guo, Lai Ye and Jinfeng Jiang
Email:{$mengshiqiao, Jiajie\_Li, jjf$}@*alumni.tongji.edu.cn*, {$weiweiguo, yelai$}@*tongji.edu.cn*
Tongji University, Shanghai, China
**denotes equal contribution* ✉*denotes corresponding author*

Network Architecture



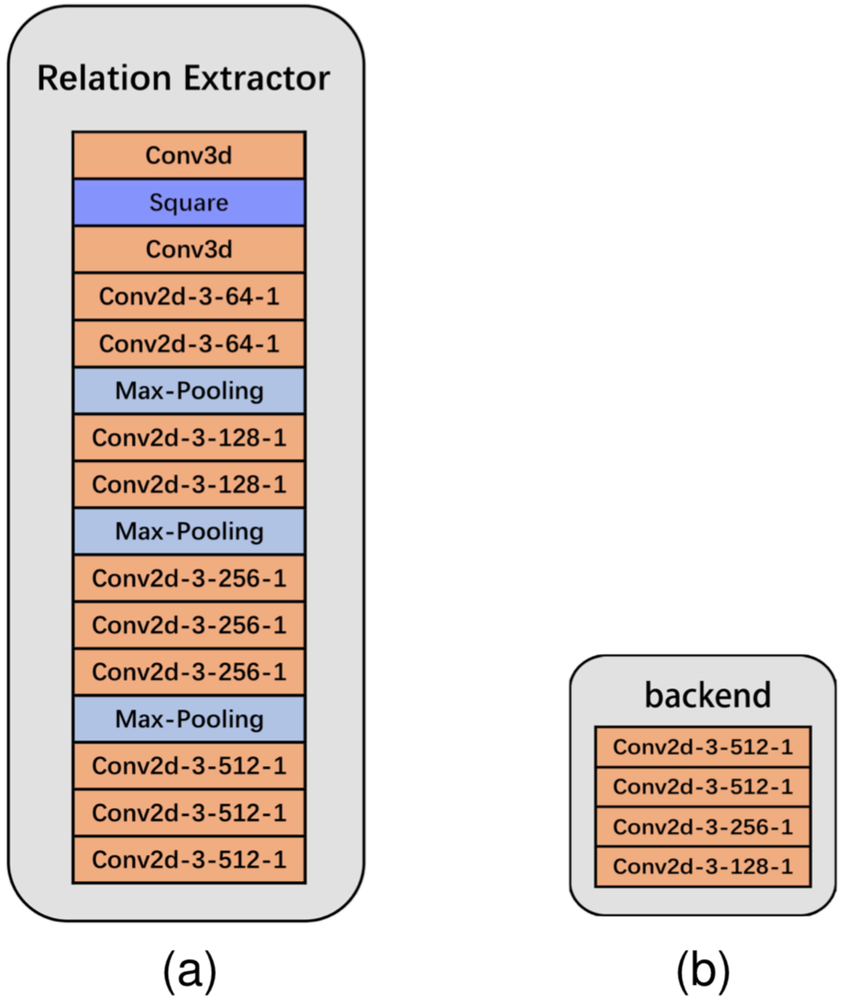
Theoretical Basis

$$R_k = \sum_{ij} \omega_{ijk} F_i^{t-1} F_j^t \tag{2}$$

$$R_k = \sum_f \omega_{kf}^R \sum_i \omega_{ik}^{t-1} F_i^{t-1} \sum_j \omega_{jk}^t F_j^t \tag{3}$$

$$\begin{aligned} R_k &= \sum_f \omega_{kf}^R (\omega_{\mathbf{ik}}^{\mathbf{t-1}^{\mathrm{T}}} F_i^{t-1} + \omega_{\mathbf{jk}}^{\mathbf{t}^{\mathrm{T}}} F_j^t)^2 \\ &= \sum_f \omega_{kf}^R \Big[2(\omega_{\mathbf{ik}}^{\mathbf{t-1}^{\mathrm{T}}} F_i^{t-1})(\omega_{\mathbf{jk}}^{\mathbf{t}^{\mathrm{T}}} F_j^t) + \\ &\quad (\omega_{\mathbf{ik}}^{\mathbf{t-1}^{\mathrm{T}}} F_i^{t-1})^2 + (\omega_{\mathbf{jk}}^{\mathbf{t}^{\mathrm{T}}} F_j^t)^2 \Big] \end{aligned} \tag{4}$$

RE(Relation Extractor)



Experiments

TABLE VII
STATISTICS OF THE FOUR DATASETS

Dataset	Resolution	Color	Total Frames	FPS	Scenes	Camera Type	Max	Min	Avg	Total	Year
UCSD [13]	158 × 238	Grey	2,000	10	1	Fixed	46	11	24.9	49,885	2008
Venice [12]	720 × 1280	RGB	167	0.5	4	Moving	421	0	215.0	35,902	2018
CrowdFlow [15]	720 × 1280	RGB	3,200	24	5	Both	911	79	319.1	1,021,064	2018
FDST [14]	1080 × 1920	RGB	15,000	30	13	Fixed	57	9	26.7	394,081	2019

* Frames:the number of video frames; Annotated FPS: frames per second of the annotated video dataset; Scenes: the number of different scenes in the dataset; MaxMin:the maximum and minimum numbers of people in the ROI of an image; Average: the average pedestrian count; Total:the total number of labeled pedestrians.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|q_i - \hat{q}_i|$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(q_i - \hat{q}_i)^2}$$

Experiments

TABLE VIII
ESTIMATION ERRORS ON THE UCSD DATASET

Method	Venue	Year	MAE	RMSE
Switch-CNN [29]	CVPR	2017	1.62	2.10
Zhang et al. [30]	CVPR	2015	1.60	3.31
ConvLSTM [8]	ICCV	2017	1.30	1.79
CSRNet [7]	CVPR	2018	1.16	1.47
Bi-ConvLSTM [8]	ICCV	2017	1.13	1.43
MCNN [6]	CVPR	2016	1.07	1.35
SANet [31]	ECCV	2018	1.02	1.29
ADCrowdNet [32]	CVPR	2019	0.98	1.25
PACNN [33]	CVPR	2019	0.89	1.18
PHNet(ours)	-	-	**0.82**	**1.05**

TABLE IX
ESTIMATION ERRORS ON THE VENICE DATASET

Method	Venue	Year	MAE	RMSE
MCNN [6]	CVPR	2016	145.4	147.3
Switch-CNN [29]	CVPR	2017	52.8	59.5
CSRNet [7]	CVPR	2018	35.8	50.0
CAN [14]	CVPR	2019	23.5	38.9
ECAN [14]	CVPR	2019	20.5	29.9
GPC [34]	IROS	2019	18.2	26.6
PHNet(ours)	-	-	**18.1**	**25.1**

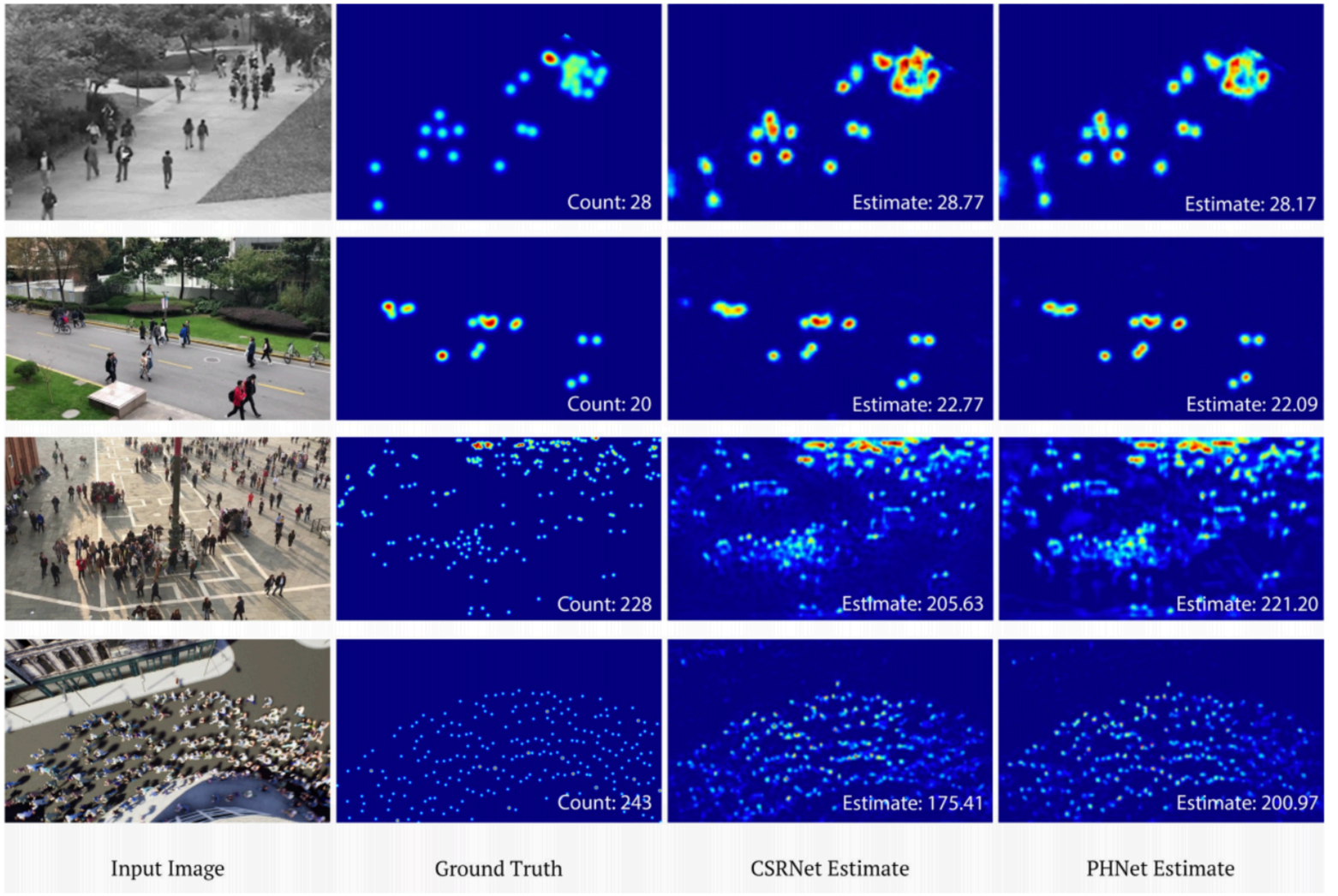
TABLE X
ESTIMATION ERRORS ON THE FDST DATASET

Method	Venue	Year	MAE	RMSE
ConvLSTM [8]	ICCV	2017	4.48	5.82
WithoutLST [24]	ICME	2019	3.87	5.16
MCNN [6]	CVPR	2016	3.77	4.88
LST [24]	ICME	2019	3.35	4.45
PHNet(ours)	-	-	**1.65**	**2.16**

TABLE XI
ESTIMATION ERRORS ON THE CROWDFLOW DATASET

Method	Venue	Year	MAE	RMSE
MCNN [6]	CVPR	2016	172.8	216.0
CSRNet [7]	CVPR	2018	137.8	181.0
CAN [14]	CVPR	2019	124.3	160.2
PHNet(ours)	-	-	**107.9**	**127.6**

Experiments



Ablation Studies

TABLE II
ABLATION STUDY FOR NUMBER OF FRAMES

Frames	MAE	RMSE
2	21.4	27.7
3	**18.1**	**25.1**
4	20.5	27.4
5	20.3	28.6
6	21.7	30.2
7	24.7	34.0
8	27.5	36.3

TABLE III
ABLATION STUDY FOR TEMPORAL MODELING METHOD

Implementation	Feature type	MAE	RMSE
Plain 3D Convolution	Additive	25.7	32.3
Relation Extractor(ours)	Multiplicative	**18.1**	**25.1**

TABLE IV
ABLATION STUDY FOR SELECTION OF SPATIAL MODEL

Spatial Model	MAE	RMSE	MAE(p)	RMSE(p)
CAN	**19.5**	**27.4**	23.5	38.9
MCNN	**31.3**	**41.4**	145.4	147.3
CSRNet	**18.1**	**25.1**	35.8	50.0

TABLE V
ABLATION STUDY FOR DIFFERENT KINDS OF INPUT DATASET

Input	MAE	RMSE
Moving dataset	**18.1**	**25.1**
Static dataset	20.3	30.3

TABLE VI
ABLATION STUDY FOR NUMBERS OF PARAMETERS AND INFERENCE TIME

Model	MAE	RMSE	C_{paras}	T_{infer}
CSRNet(ResNet50)	33.3	38.9	**26,740,546**	23.3
CSRNet(ResNet101)	31.3	37.7	45,732,609	35.1
CSRNet(ResNet152)	19.0	26.8	61,376,322	44.5
PHNet(ours)	**18.1**	**25.1**	30,094,025	**14.2**

* C_{paras}:the number of parameters; T_{infer}: time(ms) cost of inference.