PANDA: A Gigapixel-level Human-centric Video Dataset Supplementary Material

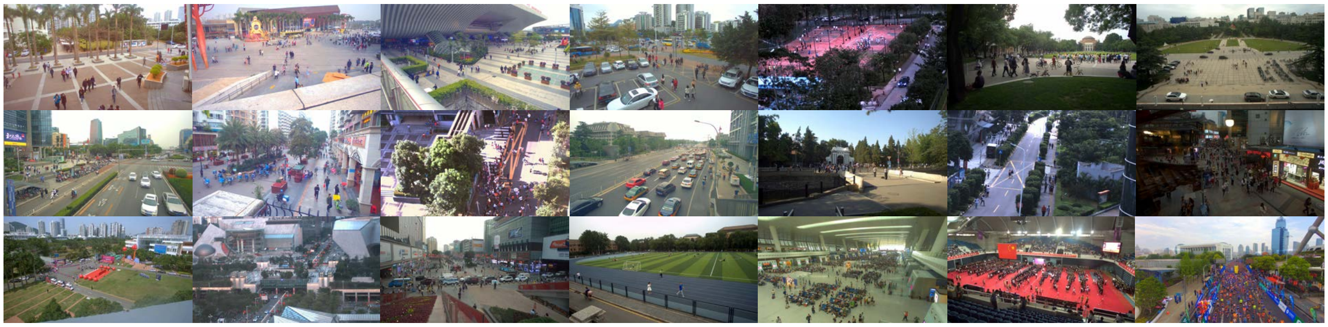

Figure 1. Overview of 21 real-world large-scale scenes in PANDA.

S. 3.1.1. Scene Display and Label Description

Currently, PANDA consists of 21 real-world large-scale scenes, as shown in Fig. 1, and the annotation details are illustrated in Tab. 1. We are continuously collecting more videos to enrich our dataset. Note that all the data was collected in public areas where photography is officially approved, and it will be published under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License [2].

S. 3.2.1. Statistical Overview of Scenes

This section includes additional statistics for each scene and training-testing set split. In Tab. 2 and Tab. 3, we give an overview of the training and testing set characteristics for PANDA and PANDA-Crowd images, respectively. In Tab. 4, we give an overview of the training and testing set characteristics for PANDA videos.

S. 4.1.1. Evaluation Metrics for Object Detection

Our evaluation metrics are the Average Precision $AP_{.50}$ and Average Recall AR, which are adopted from the MS COCO [6] benchmark. Specifically, $AP_{.50}$ is defined as the average precision at $\mathrm{IoU} = 0.50$ and AR is defined as average recall with IoU ranging in $[0.5, 0.95]$ with a stride of 0.05. To get rid of the bias towards the overcrowded frames, the maximum number of detection results on each frame is set to 500 for the calculation of AP and AR. Precision and recall is defined as follows:

$$precision = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{1}$$

$$recall = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

where TP, FP, FN are the number of True Positive, False Positive, False Negative, respectively. The Interaction-of-Union (IoU) between two bounding boxes is defined as follows:

$$\mathrm{IoU} = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

where A, B are pixel areas of the predicted and ground-truth bounding boxes respectively.

S. 4.1.2. Visualization of Object Detection Results

Fig. 2 depicts the representative failure and success cases of our detection results. As shown in the success cases, our detectors are capable to detect human body with various scale and poses by utilizing the local high-resolution visual feature. On the other hand, there are three types of failure cases: 1) confusion detection of the human-like objects; 2) duplicated detection on a single instance induced by the sliding window strategy; 3) missing detection of the human body with irregular size and scale due to occlusion or curled pose. These representative failure cases demonstrate the data diversity of our dataset that still has large room for improvement of the detection algorithms.

S. 4.2.1. Evaluation Metrics for Multiple Object Tracking

This section includes additional details regarding the definitions of the evaluation metrics for multiple objects tracking, which are partially explained in Section 4.2. The

Data	Attributes		Labels
Image	Location	Person ID	–
		Head Point	Marked in the geometric center of the human head
		Bounding Box	Estimated Full Body; Visible Body; Head
	Properties	Age	Child; Adult
		Posture	Walking; Standing; Sitting; Riding; Held in Arms
		Rider Type	Bicycle Rider; Tricycle Rider; Motorcycle Rider
		Special Cases	Fake Person; Dense Crowd; Ignore
Video	Trajectories	Person ID	–
		Bounding Box	Visible Body; Estimated Full Body (for disappearing case)
	Properties	Age	Child; Youth and Middle-aged; Elderly
		Gender	Male; Female
		Face Orientation	↑↓→←↗↘↙↖
		Occlusion Degree	W/O Occlusion; Partial Occlusion; Heavy Occlusion; Disappearing
	Group	Group ID	–
		Intimacy	Low; Middle; High
		Group Type	Acquaintance; Family; Business
	Interaction	Begin/End Frame	–
		Interaction Type	Physical Contact; Body Language; Face Expressions; Eye Contact; Talking
		Confidence Score	Low; Middle; High

Table 1. Annotation Details in PANDA dataset.

Scene	#Sub-scene	#Image	Resolution	Mean #Person	Mean #Special Case	Mean Person Height	Mean Occlusion Ratio	Camera Height
Training Set								
University Canteen	1	30	26753×15052	52.7	23.7	906.79	0.11	2nd Floor
Xili Crossroad	1	30	26753×15052	174.2	27.3	506.78	0.13	2nd Floor
Train Station Square	2	15/15	26583×14957	272.1	75.8	328.05	0.11	2nd Floor
Grant Hall	1	30	25306×14238	133.1	22.6	583.38	0.13	1st Floor
University Gate	1	30	26583×14957	122.6	43.9	617.88	0.20	1st Floor
University Campus	1	30	26088×14678	223.0	26.7	293.80	0.08	8th Floor
East Gate	1	30	25831×14533	175.7	37.4	201.43	0.14	2nd Floor
Dongmen Street	1	30	25151×14151	289.4	79.4	551.16	0.15	2nd Floor
Electronic Market	1	30	25306×14238	571.6	113.4	339.17	0.23	2nd Floor
Ceremony	1	30	25831×14533	250.3	51.3	308.69	0.11	5th Floor
Shenzhen Library	2	15/15	32129×24096 / 31746×23810	190.9	59.1	321.77	0.13	20th Floor
Basketball Court	2	15/15	31753×23810 / 31746×23810	86.7	10.4	928.29	0.07	10th Floor
University Playground	2	15/15	27098×15246 / 25654×14434	127.5	14.4	307.45	0.04	2nd Floor
Testing Set								
OCT Habour	1	30	26753×15052	278.8	48.5	495.34	0.10	2nd Floor
Nanshani Park	1	30	32609×24457	83.6	24.9	1,108.77	0.14	5th Floor
Primary School	2	15/15	31760×23810	233.9	24.0	1,096.56	0.08	19th Floor
New Zhongguan	1	30	26583×14957	352.6	85.2	353.08	0.16	2nd Floor
Xili Street	2	30/15	26583×14957 / 26753×15052	118.4	47.2	642.51	0.13	2nd Floor

Table 2. Statistics and train-test set split for 18 scenes of PANDA images. '#' represents 'The number of'; Sub-scene represents data captured in the same scene, but with different viewpoints or recording time; 'Mean' represents the mean of the value for each image; Person height is calculated in pixels; Occlusion Ratio is the ratio of the visible body bbox area to the estimated full body bbox area.

Scene	#Image	Resolution	Mean #Person	Camera Height
Training Set				
Marathon	15	26908×15024	3,619.2	4th Floor
Graduation Ceremony	15	26583×14957	1,483.0	2nd Floor
Testing Set				
Waiting Hall	15	26558×14828	3,039.1	2nd Floor

Table 3. Statistics and train-test set split for 3 scenes of PANDA-Crowd images. '#' represents 'The number of'; 'Mean' represents the mean of the value for each image.

Scene	#Frame	FPS	Resolution	#Tracks	#Boxes	#Groups	#Single Person	Camera Height
Training Set								
University Canteen	3,500	30	26753×15052	295	335.2k	75	123	2nd Floor
OCT Habour	3,500	30	26753×15052	736	1,270.1k	205	191	2nd Floor
Xili Crossroad	3,500	30	26753×15052	763	1,065.0k	163	393	2nd Floor
Primary School	889	12	34682×26012	718	465.6k	117	119	19th Floor
Basketball Court	798	12	31746×23810	208	118.4k	34	54	10th Floor
Xinzhongguan	3,331	30	26583×14957	1,266	1,626.0k	186	857	2nd Floor
University Campus	2,686	30	25479×14335	420	658.6k	83	123	8th Floor
Xili Street 1	3,500	30	26583×14957	662	950.0k	144	325	2nd Floor
Xili Street 2	3,500	30	26583×14957	290	425.7k	59	152	2nd Floor
Huaqiangbei	3,500	30	25306×14238	2,412	3,054.5k	310	1,730	2nd Floor
Testing Set								
Train Station Square	3,500	30	26583×14957	1,609	1,682.7k	178	1,213	2nd Floor
Nanshan i Park	889	12	32609×24457	402	132.6k	78	199	5th Floor
University Playground	3,560	30	25654×14434	309	574.3k	60	165	2nd Floor
Ceremony	3,500	30	25831×14533	677	1,444.7k	143	317	5th Floor
Dongmen Street	3,500	30	26583×14957	1,922	1,676.4k	331	1,170	2nd Floor

Table 4. Statistics and train-test set split for 15 scenes of PANDA videos. '#' represents 'The number of'; FPS represents 'Frames Per Second'.

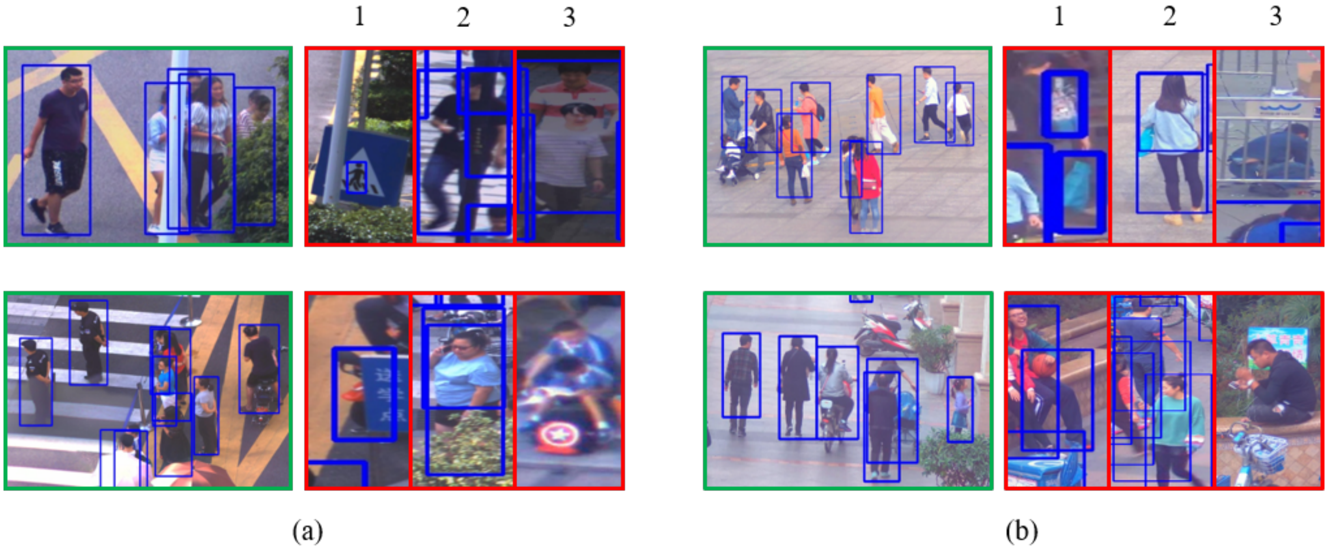

Figure 2. Success cases (green) and failure cases (red). (a) Cascade R-CNN on Full Body. (b) Faster R-CNN on Visible Body. The failure cases can be summarized into three types: (1) confusion detection of the human-like objects; (2) duplicated detection on a single instance induced by the sliding window strategy; (3) missing detection of the human body with irregular size and scale due to occlusion or curled pose.

measurements are adopted from the MOT Challenge [7] benchmarks. In MOT Challenge, 2 sets of measures are employed: The CLEAR metrics proposed by [8], and a set of track quality measures introduced by [9].

The distance measure, *i.e.*, how close a tracker hypothesis is to the actual target, is determined by the intersection over union (IoU) between estimated bounding boxes and the ground truths. The similarity threshold t_d for true positives is empirically set to 50%.

The Multiple Object Tracking Accuracy (MOTA) combines three sources of errors to evaluate a tracker's performance, defined as

$$\mathrm{MOTA} = 1 - \frac{\sum_t (\mathrm{FN}_t + \mathrm{FP}_t + \mathrm{IDSW}_t)}{\sum_t \mathrm{GT}_t} \tag{4}$$

where t is the frame index. FN, FP, IDSW and GT respectively denote the numbers of false negatives, false positives, identity switches and ground truths. The range of MOTA is $(-\infty, 1]$, which becomes negative when the number of errors exceeds the ground truth objects.

Multiple Object Tracking Precision (MOTP) is used to measure misalignment between annotated and predicted object locations, defined as

$$\text{MOTP} = 1 - \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t} \tag{5}$$

where c_t denotes the number of matches in frame t and $d_{t,i}$ is the bounding box overlap of target i with its assigned ground truth object. MOTP thereby gives the average overlap between all correctly matched hypotheses and their respective objects and ranges between $t_d := 50\%$ and 100%. According to [7], in practice, it mostly quantifies the localization accuracy of the detector, and therefore, it provides little information about the actual performance of the tracker.

S. 4.3.1. Network Structure

Global Trajectory. To obtain the global trajectory edge set E_{global} and edge weight function $w_{global} : E_{global} \rightarrow \mathbb{R}$, we use a simple LSTM(4 layers,128 hidden state) and embedding learning with triplet loss(margin=0.5) to extract the sequence embedding vector for each vertex v(denoted as $F_v \in \mathbb{R}^{512}, v \in V$). And then the edge weight function is calculated by:

$$w_{global}(e) = ||F_u - F_v||_2, where\ e = \{u, v\}\ and\ u, v \in V \tag{6}$$

More specifically about embedding network, the input trajectory is the variable-length sequence where each element $\in \mathbb{R}^6$ consists of bounding box coordinates(4 scalar), face orientation angle(1 scalar,optional), and timestamp(1 scalar). The output F_v is obtained by concatenating the hidden state vector and cell output vector in LSTM. The supervision signal is given by triplet loss which enforces trajectories from the same group to have small L2 distance in embedding feature space and trajectories from different groups to have a large distance.

Local Interaction. As mentioned in the paper, calculating interaction score for each pair of human entities is inefficient and we only check a subset of entity pairs. In other words, given that $E_{local} \subset E_{global}, |E_{local}| << |E_{global}|$, $w_{local} : E_{local} \rightarrow \mathbb{R}$ is the target. More specifically, for each $e \in E_{local}$, several local video candidate clips $clip_e = \{clip_{e,i}\}$ is firstly cropped spatially and temporally from full video by filtering using the relative distance between 2 entities which is possible for interaction. The $clip_{e,i}$ is the variable-length sequence where each frame$\in \mathbb{R}^{4 \times H \times W}$ consists of 3 channel RGB image and 1 channel interaction persons mask.And then for each $clip_{e,i}$ we use Spatial-temporal 3D ConvNet[3] as local video classifier which estimates the interaction score. Finally, w_{local} is obtained by averaging interaction score of all the $clip_e$ as follow:

$$w_{local}(e) = \frac{\sum_{i=0}^{\#of clips}(ConvNet(clip_{e,i}))}{\#of clips}, e \in E_{local} \tag{7}$$

We use pre-trained weight from large scale dataset Kinetics[4] and follows the same hyper-parameter, loss function as [3].

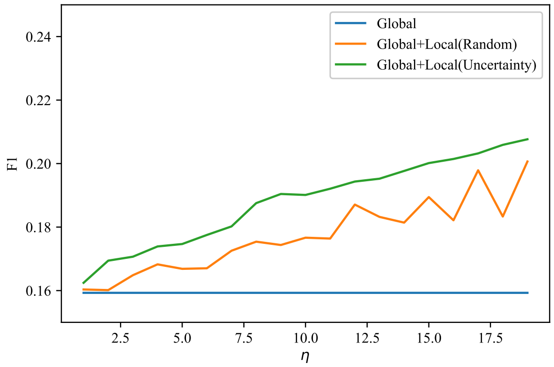

Figure 3. Sensitivity study of η(average on τ). As η increase, performances of all three model are improved and computation consumption increase as well. However, using global feature, local feature and uncertainty can achieve higher performance than random zoom in policy or without local feature under different η value.

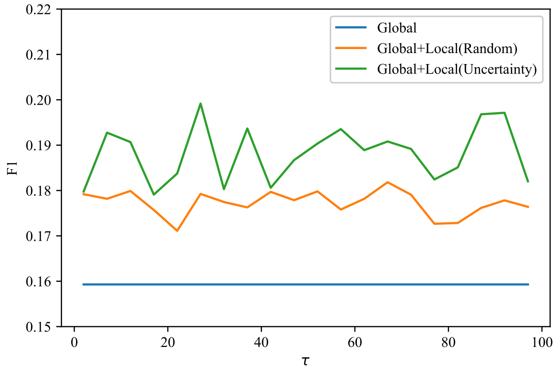

Figure 4. Sensitivity study of τ(average on η). Using global feature, local feature and uncertainty can achieve higher performance than random policy or without local feature. And there is no significant increase in performance as tau increasing from 10 to 510.

S. 4.3.2. Zoom-in policy

Zoom-in module solves the problem of selecting a subset of edge$E_{uncertainty}$ to calculate local interaction score given E_{global}. And each edge $e \in E_{uncertainty}$ is further fed into the local interaction module and then E_{local}, w_{local} are obtained as above. There are 2 methods compared in the paper: random selection and uncertainty-based method. For the former one, $E_{uncertainty}$ consists of η samples which are randomly selected from E_{global} and predicted to be positive. For the latter one, the top η positive predicted uncertain edges are selected. To estimate the uncertainty, stochastic dropout sampling[5] is adopted. More specifically, with dropout layer activated and perform inference τ times per input. Thus for each edge score there are τ estimations and we can use the variance among the estimations as the desired uncertainty.

S. 4.3.3. Edge Merging Strategy

Given E_{global}, E_{local} and w_{global}, w_{local} defined on them, label propagation strategy[10] is adopted to delete or merge some edges with adaptive threshold in a iterative manner. While edges are gradually deleted, the graph is divided into several disconnected components which is the group detection result.

S. 4.3.4. Trajectory source in group detection

We encourage users to explore the integrated solution which takes MOT result trajectory as group detection input. However, in our experiment, even the SOTA MOT method can not address the serious ID-switch, trajectory fragmentation problem. Thus, we separate the MOT task and group detection task for the first step benchmark and the previous incremental effectiveness experiment. The released dataset provides sufficient annotation and we encourage users to explore the more robust MOT methods or the integrated solution of 2 tasks. As a result of using trajectory annotation, the training-testing set split is different from previous task. In the group detection task, we use Training set in Tab. 4 to train and test. More specifically, scene *University Canteen* is used as the testing set and the rest 8 scenes are used as training sets.

S. 4.3.5. Evaluation Metrics

As discussed in [1], the half metric refers to a single detected group prediction that is positive if the detected group contains at least half of the elements of the Ground Truth group (and vice-versa). And then we can calculate precision, recall, and F1 based on the positive and negative samples. More specifically, each detected group (Grp_{pd}) as well as ground truth(Grp_{gt}) is a set of group member:

$$Grp_* = \{v|v \in V \ and \ v \ belong \ to \ the \ group\} \tag{8}$$

And one detected group is regarded as correct under half metric if and only if it satisfy the following:

$$\frac{Grp_{pd} \cap Grp_{gt}}{max(|Grp_{pd}|, |Grp_{gt}|)} > 0.5 \tag{9}$$

References

[1] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Discovering groups of people in images. In *European conference on computer vision*, pages 417–433. Springer, 2014. 5

[2] Creative Commons. Commons attribution-noncommercial-sharealike 4.0 license. https://creativecommons.org/licenses/by-nc-sa/4.0/. 1

[3] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 4

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 4

[5] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015. 5

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[7] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 3, 4

[8] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, John Garofolo, Djamel Mostefa, and Padmanabhan Soundararajan. The clear 2006 evaluation. In *International evaluation workshop on classification of events, activities and relationships*, pages 1–44. Springer, 2006. 3

[9] Bo Wu and Ram Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 951–958. IEEE, 2006. 3

[10] Xiaohang Zhan, Ziwei Liu, Junjie Yan, Dahua Lin, and Chen Change Loy. Consensus-driven propagation in massive unlabeled data for face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 568–583, 2018. 5