Reconstructing Groups of People with Hypergraph Relational Reasoning —Supplementary Material—

Buzhen Huang[1] Jingyi Ju[1] Zhihao Li[2] Yangang Wang[1*]

[1]Southeast University, China
[2]Huawei Noah's Ark Lab

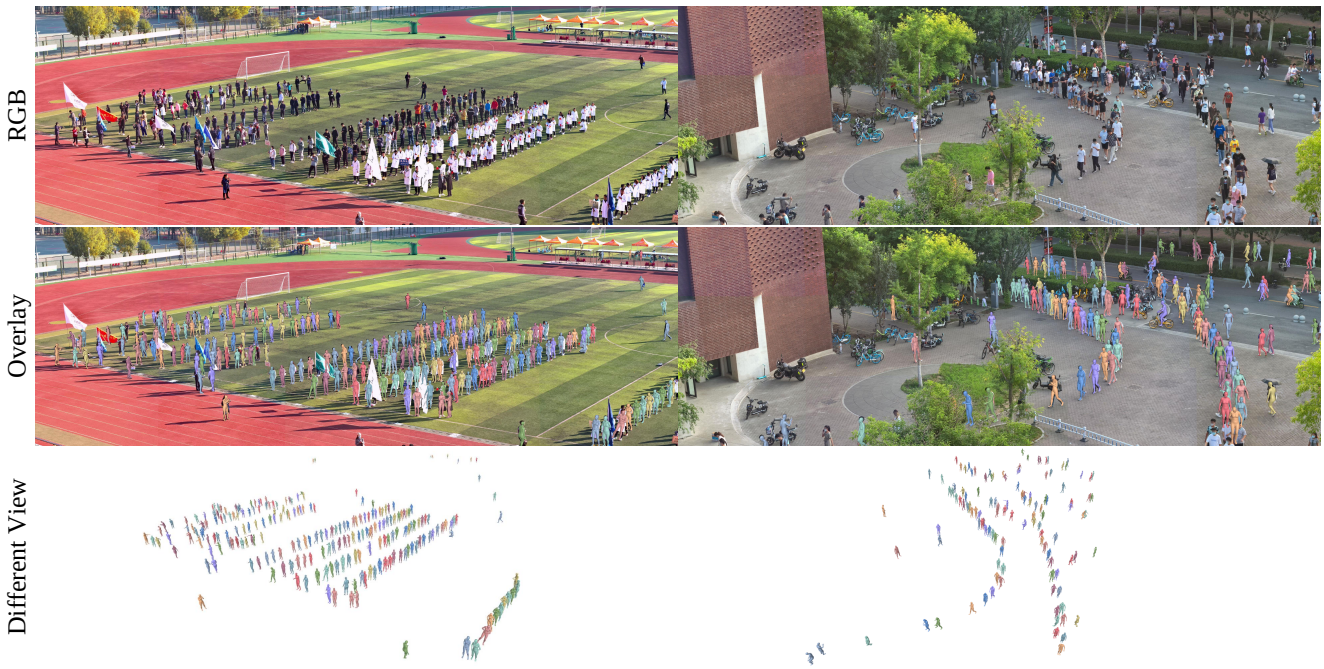

Figure 1: Our method can produce accurate body meshes with reasonable spatial positions from monocular large-scale images.

1. Introduction

In the supplementary material, we first provide the schematic, pseudocode, and more details to help the readers to understand the hypergraph relational reasoning (Sec. 2). Then, we introduce the procedures to produce pseudo ground-truth (Sec. 3). The implementation details (Sec. 4) and training data (Sec. 6) are also described to help the reproduction of the experimental results. Finally, more comparisons, analyses, and qualitative experiments are conducted to further demonstrate the superiority of the proposed method (Sec. 7).

2. Relational reasoning

We provide a schematic (Fig. 2) and pseudocode (Algorithm 1) of the relational reasoning to promote the readers to understand the procedures. To form the hypergraphs, we first initialize all nodes with the extracted individual features. Then, the human features are used to calculate the affinity map for inferring hyperedges. Since the human groups are unordered structures, which cannot be defined in advance, we employ a greedy algorithm approximation to predict the connection relationships of hyperedges in different scales. Specifically, we first select a node v_i and then add new nodes that have maximum affinity values with v_i. The nodes on the same hyperedge will be regarded as a

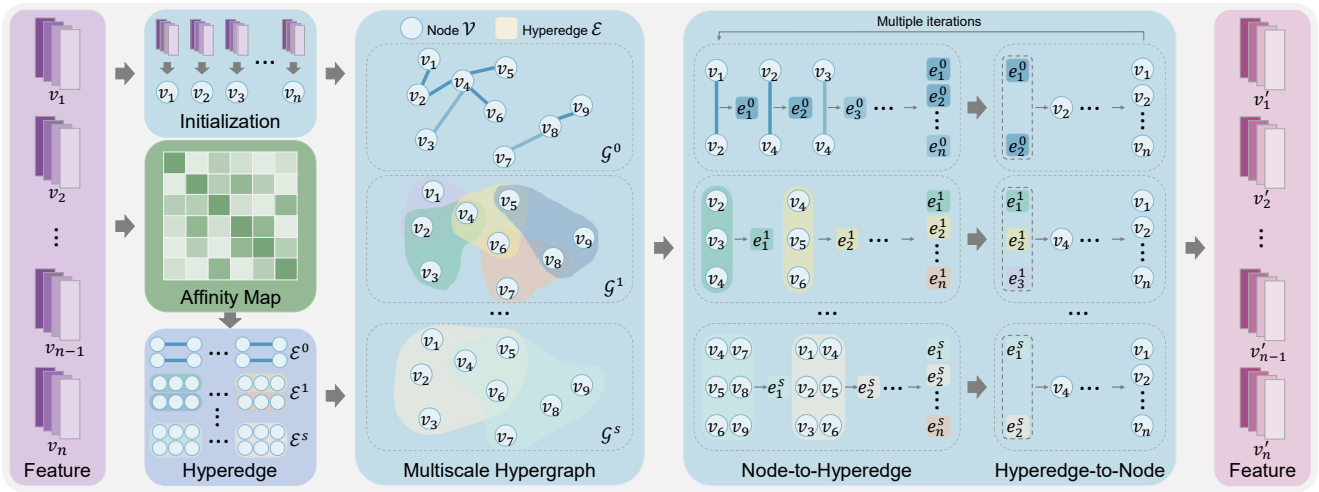

Figure 2: Schematic of the relational reasoning. We initialize the nodes and infer the hyperedges with the extracted individual features. After the hypergraphs are formed, we conduct node-to-hyperedge and hyperedge-to-node phases to achieve the relational reasoning.

group. After the hypergraphs are formed, we conduct node-to-hyperedge and hyperedge-to-node phases to achieve the relational reasoning. As shown in the Fig. 2, the node features are first aggregated to hyperedges. Then, we update the node features with all associated hyperedges. The two phases are executed for several iterations. Finally, the updated node features are output for the regression. The execution steps can be found in Algorithm 1.

3. Pseudo ground-truth annotator

In this section, we introduce the detailed procedures of our pseudo ground-truth annotator. The procedures are similar to EFT [8], which adapts 3D human models to ground-truth 2D poses with a strong pose prior. The critical difference between our annotator and EFT is that we consider the multi-person relationships and constraints. Thus, our annotator can obtain more reasonable spatial distributions and ordinal relationships. Specifically, we first train our model on common crowd data. Due to the domain gap between common and large-scale scenarios, the network may not predict accurate results when applied to large images. Thus, we finetune the network parameters on each image with the following constraints:

$$\mathcal{L}_{pseudo} = \mathcal{L}_{\text{reproj}} + \mathcal{L}_{\text{crowd}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{prior}}. \quad (1)$$

$\mathcal{L}_{\text{reproj}}$ and $\mathcal{L}_{\text{crowd}}$ are the same as main manuscript. A depth ordering-aware loss [6] is also used to ensure the correct ordinal relations for close people.

$$L_{\text{depth}} = \sum_{u \in \mathcal{S}} \log\left(1 + \exp\left(D_{y(u)}(u) - D_{\hat{y}(u)}(u)\right)\right). \quad (2)$$

The loss function penalizes inconsistent depth ordering assisted by a differentiable depth renderer [9]. $y(u)$ and $\hat{y}(u)$ are ground-truth and predicted person index at pixel location u. $D_{y(u)}(u)$ is the depth value at u for $y(u)$th person, and S is the set of pixels that have conflicting depth ordering. To query ground-truth depth value, we predict the instance segmentation with Pose2seg [22]. More details can refer to [6]. Since this constraint has low efficiency for crowd images, we only apply it in the pseudo-ground truth annotator and do not use it in the network training.

A regularization term is also used to prevent overfitting:

$$\mathcal{L}_{\text{prior}} = \frac{1}{N}\sum_{n=1}^{N} \|[\beta^n, \theta^n] - [\beta^{\hat{n}}_{init}, \theta^{\hat{n}}_{init}]\|_2^2, \quad (3)$$

where β_{init} and θ_{init} are initial values of the first prediction. We optimize the network parameters for several iterations based on the above constraints and finally output the estimated results as the pseudo ground-truth.

4. Implementation details

We implement the network with PyTorch [17]. The backbone network is a ResNet-50 [4], which encodes the human image patch to a feature vector. We pretrain the backbone network with a single-person mesh recovery task, and then use it to form the overall framework. To consider both common and crowded scenes, the relational reasoning network is designed to have 11 nodes. When the number of people in the image is less than the number of nodes, we fill the empty node with 0 and mask it in the relational reasoning. For the images with known camera parameters, we use the ground-truth focal length in the training and inference; otherwise, we use an approximate value of $\sqrt{w^2 + h^2}$, where w and h are the width and height of the image. The network is trained on a desktop with an Intel(R) Core(TM) i9-11900F CPU and a GPU of NVIDIA GeForce

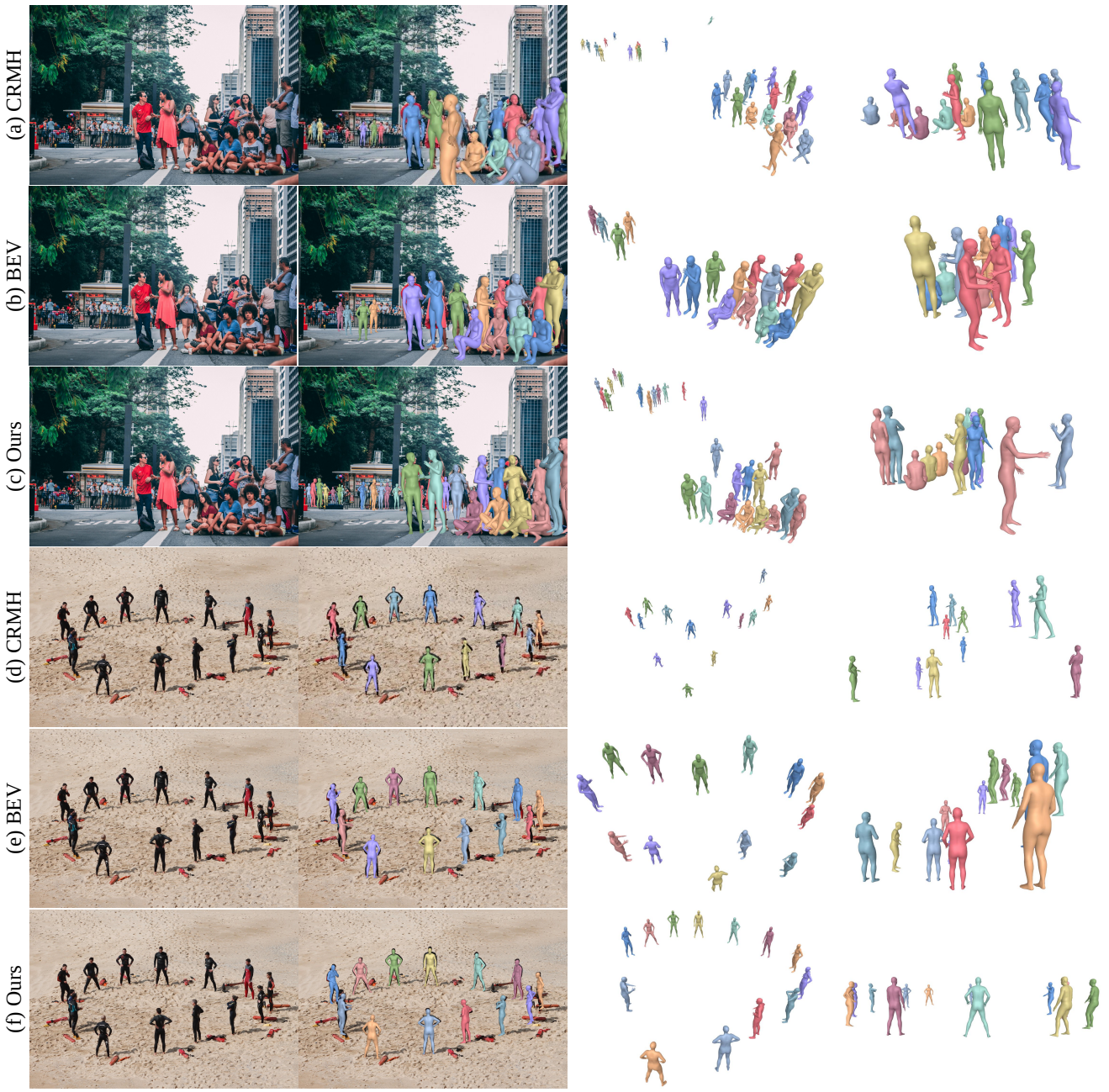

Figure 3: Comparison with BEV [19] and CRMH [6]. Our method regresses human meshes with more accurate spatial positions.

RTX 3090, using the AdamW optimizer [14] with a learning rate of 1e-4 and a batch size of 32. To obtain pseudo ground-truth annotations, we run 260 iterations to adapt the pretrained model to in-the-wild images with a learning rate of 1e-5.

5. Metrics

We describe the details of the metrics used in large-scale scenes. The Procrustes-aligned pair-wise percentual distance similarity (PA-PPDS) [1] measures the location distribution of the crowd after applying a Procrustes Analysis between the estimated crowd and the real crowd.

$$PA-PPDS = \frac{\sum_{k=1}^{N-1}\sum_{i=k+1}^{N} 1-\min\left(d_{ik},1\right)}{C_N^2}, \qquad d_{ik} = \left|\frac{\|E_k-E_i\|-\|G_k-G_i\|}{\|G_k-G_i\|}\right|, \tag{4}$$

Testing set	Panoptic	GigaCrowd	JTA
Training set	Human3.6M MuCo-3DHP MSCOCO MPII CrowdPose	Human3.6M MuCo-3DHP MSCOCO MPII CrowdPose Panda	JTA

Table 1: The benchmarks and corresponding training data.

where N is the number of people in the image, and E_i and G_i represent the estimated and ground-truth locations of the ith person, respectively. Since a low detection threshold may result in redundant reconstruction, a redundant punishment (RP) [1] is also used to evaluate the redundancy.

$$\mathrm{RP} = \min\left(1, \max\left(\frac{N_{\text{pred}}}{N_{gt}} - \text{tre}, 0\right)\right) \tag{5}$$

where N_{pred} N_{gt} are the number of persons in the prediction and ground truth, respectively. tre $= 1.02$ is a threshold.

6. Dataset

Panda [20] is the first giga-pixel level video dataset captured in real-world scenarios. This dataset contains crowded scenes with a large field of view and high-resolution details. However, only the bounding-boxes in this dataset are available. We annotate the ground-truth 2D poses according to the bounding-boxes. Based on the 2D poses, we further build pseudo ground-truth 3D crowd for this dataset and use the generated annotations to train our model.

GigaCrowd [1] is a giga-pixel level image dataset, which captures several large-scale crowded scenarios in the real world. Since it is a dataset for a competition, only the ground-truth 3D root positions and 2D poses in its training set are publicly available. Thus, we conduct quantitative comparisons on its training set. In addition, the testing set is also used to conduct the qualitative experiments.

JTA [3] is a synthetic dataset for human pose estimation in large urban scenarios, which is collected from the realistic video-game the Grand Theft Auto V. Since the dataset only has 3D joint position annotations, we fit the SMPL model to the 3D poses to get human meshes for training. The evaluation is conducted on the standard testing set to demonstrate the effectiveness of our model in large-scale crowded scenes.

Panoptic [7] has several people captured in a controlled scene, which is used for evaluation only. This dataset is challenging in terms of complex interactions and difficult camera viewpoints. We follow the protocol used in [21] to evaluate our model.

Human3.6M [5] is a single-person indoor dataset that contains 11 professional actors in 17 scenarios. We follow the previous work [19] to use the subjects S1, S5, S6, S7 and S8 with Mosh [13] annotations for training.

Iterations	1	2	3	4	5
MPJPE	106.9	106.7	106.6	106.8	107.0

Table 2: Ablation on the number of iterations on Panoptic. Different iterations achieve very similar performance.

Individual	Transformer	hypergraph-(1)	hypergraph-(1,3,5)
129.2	112.8	113.4	106.6

Table 3: Ablation studies on Panoptic dataset. "Individual" removes the relational reasoning. "Transformer" uses a transformer-based network for relational reasoning. "(1,3,5)" means 3 scales with group sizes of 1, 3, and 5. The numbers are MPJPE.

MuCo-3DHP [15] is a synthetic multi-person 3D dataset. It composites images of single-person with 3D pose ground truth from the existing MPI-INF-3DHP dataset. We use the same version with 3DMPPE [16] for training.

MSCOCO [12], MPII [2], CrowdPose [10] are in-the-wild 2D pose datasets. We use pseudo ground-truth from [11] and our work for the training.

MuPoTS [15] contains ground-truth 3D poses for up to 3 subjects. It is captured with a multi-view markerless system. We use it to conduct qualitative comparisons.

AGORA [18] is a synthetic multi-person dataset with absolute human mesh annotations. It uses textured human scans in diverse poses and clothes to build the dataset. Each image has 5-15 people with various occlusions. We use AGORA to evaluate the method on occlusion cases.

7. Extended experiments

7.1. Qualitative comparison to SOTA methods

In Fig. 3, we further conduct several comparisons to SOTA multi-person mesh recovery methods on Internet images in common scenes. In Fig. 3 (b), we found that BEV cannot recover people in the distance due to the scale variations. Although BEV and CRMH can produce plausible results in the camera perspective, their spatial positions may be incorrect. In contrast, our method can regress accurate meshes with reasonable spatial distribution with the relational reasoning and crowd constraints. In addition, 3DCrowdNet and ROMP are recent works that do not generate absolute positions. We also compared our method with these baseline methods in Fig. 4. The results on Panoptic, MuPoTS, JTA, and AGORA datasets show that our method is more robust to scale variations, mutual occlusions, and truncations. We further show more results on GigaCrowd dataset in Fig. 1. Our method works well on extremely large scenes with crowded people.

7.2. Collectiveness

We visualize the human feature affinity maps $\mathcal{A}$ in Fig. 6 for correlated and unrelated groups. Based on the human correlations, we apply an optimization in the group infer-

Figure 4: Comparison with CRMH, BEV, 3DCrowdNet, and ROMP on Panoptic, MuPoTS, JTA, and AGORA. Our method is more robust to scale variations and mutual occlusions.

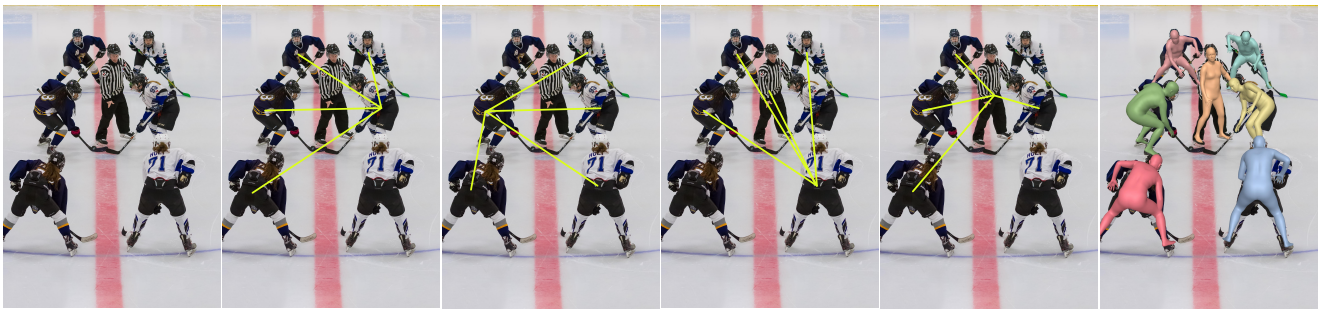

Figure 5: Visualization of different groups. The group with higher pose similarity produces a higher collectiveness factor.

ence stage to form the groups for the relational reasoning. The adjacency matrices, $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$, show that the people with a high pose similarity are assigned to the same group. We also show the formed groups in Fig. 5. The yellow lines denote the generated groups with a size of 5. Thus, the group information can be exploited to improve the

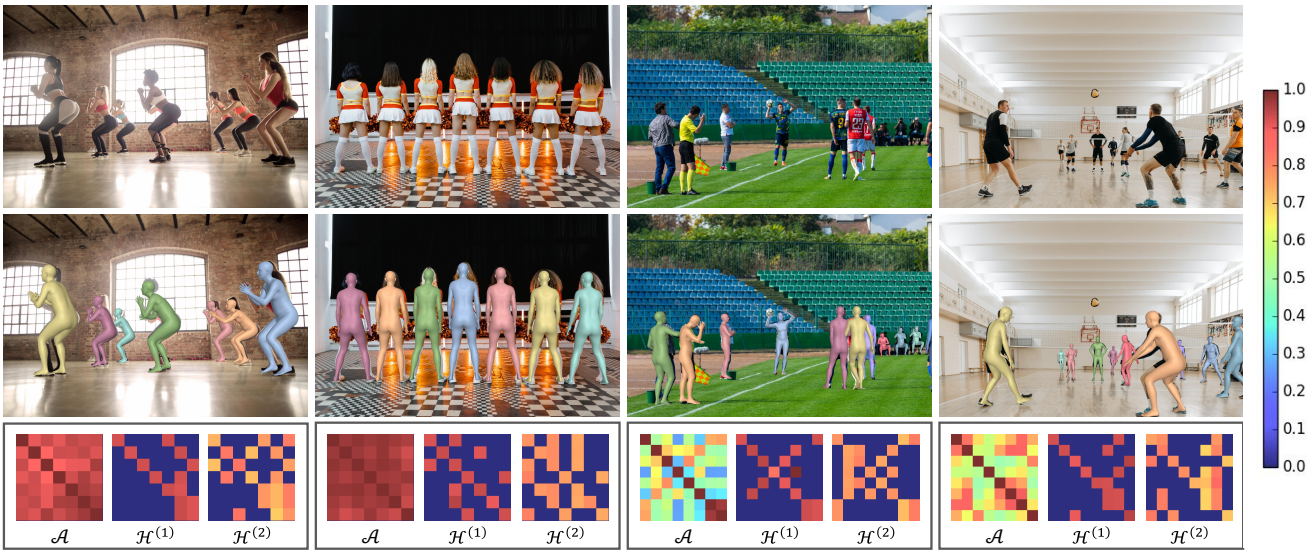

Figure 6: We show the affinity maps $\mathcal{A}$ and adjacency matrices $\mathcal{H}^{(s)}$ for correlated (column 1 and 2) and unrelated groups (column 3 and 4). The color in $\mathcal{A}$ and $\mathcal{H}^{(s)}$ means the correlation and collectiveness factor, respectively.

reconstruction for each individual. In addition, we calculate the effectiveness factor in the node-to-hyperedge phase and demonstrate that the group with higher pose similarity can lead to a higher effectiveness factor. With the group features, our model can regress more accurate body meshes and precise ordinal relationships.

7.3. The impact of iterations

In Tab. 2, we study the effect of different iteration numbers on the reconstruction performance. The experiment is conducted on Panoptic. The results show that the number of iterations has a slight effect on the estimation, and a moderate iteration number achieves the best results.

7.4. Extended ablations on Panoptic

Since GigaCrowd does not contain 3D pose annotations, we further investigate the effectiveness of our method on Panoptic by measuring the reconstructed 3D joint positions in Tab. 3. Panoptic has collective motions and complex mutual occlusions, which is also a useful tool to validate our design. The results show that group features are essential in this situation. We can reconstruct more accurate 3D poses with the relational reasoning.

7.5. Failure cases

We also show some failure cases in Fig. 7 to further discuss the limitations. Since the framework still adopts a top-down strategy, the strongly ambiguous pixel-level image features may confuse the network prediction. For severe mutual occlusions, our method cannot produce plausible results. The problem could be solved by incorporating 2D semantics or pose prior knowledge. Our method also

Figure 7: Failure cases. Our method still cannot address severe pixel-level ambiguities. In addition, in cases where people are in different planes, a too-large crowd loss weight may result in wrong absolute positions.

cannot estimate accurate 3D models for kids. Directly using SMPL to represent kids may produce strange body shapes and wrong spatial positions. Besides, the current crowd constraint is not generalized for all scenarios. As shown in the auditorium case in Fig. 7, although the overlay image seems correct, the distances of people in different rows are large. It is induced by a too-large crowd loss weight, and all people are dragged to the same plane. Future works could integrate the scene semantics to address this limitation.

Algorithm 1 Pseudocode of hypergraph relational reasoning.

Input: individual features $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$; number of iterations T; number of scales S; group size K
Output: updated individual features $\mathcal{V}' = \{v'_1, v'_2, \cdots, v'_N\}$

initialize output features $D = \{\}$
calculate affinity map $\mathcal{A}$
for $s = 1$ to S **do**
  initialize node $\mathcal{V}^{(s)}$ = $\mathcal{V}$
  infer adjacency matrix $\mathcal{H}^{(s)}$ with $\mathcal{A}$ and $K^{(s)}$
  node-to-hyperedge phase $\mathcal{E}^{(s)}$ = $MLP(\mathcal{V}^{(s)}, \mathcal{H}^{(s)})$
  for $i = 1$ to $T - 1$ **do**
    hyperedge-to-node phase $\mathcal{V}^{(s)}$ = $MLP(\mathcal{V}^{(s)}, \mathcal{E}^{(s)}, \mathcal{H}^{(s)})$
    node-to-hyperedge phase $\mathcal{E}^{(s)}$ = $MLP(\mathcal{V}^{(s)}, \mathcal{H}^{(s)})$
    $i = i + 1$
  end for
  hyperedge-to-node phase $\mathcal{V}^{(s)\prime}$ = $MLP(\mathcal{V}^{(s)}, \mathcal{E}^{(s)}, \mathcal{H}^{(s)})$
  collect features at scale s $D \leftarrow \mathcal{V}^{(s)\prime}$
  $s = s + 1$
end for
get updated individual features $\mathcal{V}' \leftarrow D$

References

[1] Gigacrowd challenge. https://www.gigavision.cn/track/track/?nav=GigaCrowd. 3, 4

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014. 4

[3] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *ECCV*, pages 430–446, 2018. 4

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2013. 4

[6] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, pages 5579–5588, 2020. 2, 3

[7] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, pages 3334–3342, 2015. 4

[8] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *3DV*, pages 42–52, 2021. 2

[9] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, pages 3907–3916, 2018. 2

[10] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, pages 10863–10872, 2019. 4

[11] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 4

[12] 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 *ECCV*, pages 740–755, 2014. 4

[13] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. *TOG*, 33(6):1–13, 2014. 4

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 3

[15] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, pages 120–130, 2018. 4

[16] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *ICCV*, pages 10133–10142, 2019. 4

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 2

[18] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora:

Avatars in geography optimized for regression analysis. In *CVPR*, pages 13468–13478, 2021. 4

[19] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, pages 13243–13252, 2022. 3, 4

[20] Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, et al. Panda: A gigapixel-level human-centric video dataset. In *CVPR*, pages 3268–3278, 2020. 4

[21] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, pages 2148–2157, 2018. 4

[22] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *CVPR*, pages 889–898, 2019. 2