Dynamic Graph Reasoning for Multi-person 3D Pose Estimation

Zhongwei Qiu*
University of Science and Technology Beijing
qizhongwei@xs.ustb.edu.cn

Jian Wang
Baidu
wangjian33@baidu.com

Qiansheng Yang*
Baidu
yangqiansheng@baidu.com

Dongmei Fu
University of Science and Technology Beijing
fdm_ustb@ustb.edu.cn

ABSTRACT
Multi-person 3D pose estimation is a challenging task because of occlusion and depth ambiguity, especially in the cases of crowd scenes. To solve these problems, most existing methods explore modeling body context cues by enhancing feature representation with graph neural networks or adding structural constraints. However, these methods are not robust for their single-root formulation that decoding 3D poses from a root node with a pre-defined graph. In this paper, we propose GR-M3D, which models the Multi-person 3D pose estimation with dynamic Graph Reasoning. The decoding graph in GR-M3D is predicted instead of pre-defined. In particular, it firstly generates several data maps and enhances them with a scale and depth aware refinement module (SDAR). Then multiple root keypoints and dense decoding paths for each person are estimated from these data maps. Based on them, dynamic decoding graphs are built by assigning path weights to the decoding paths, while the path weights are inferred from those enhanced data maps. And this process is named dynamic graph reasoning (DGR). Finally, the 3D poses are decoded according to dynamic decoding graphs for each detected person. GR-M3D can adjust the structure of the decoding graph implicitly by adopting soft path weights according to input data, which makes the decoding graphs be adaptive to different input persons to the best extent and more capable of handling occlusion and depth ambiguity than previous methods. We empirically show that the proposed bottom-up approach even outperforms top-down methods and achieves state-of-the-art results on three 3D pose datasets.

CCS CONCEPTS
• Computing methodologies → Object recognition.

KEYWORDS
Human Pose Estimation, Multi-person, Graph Reasoning

*Equal Contribution

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:
Zhongwei Qiu, Qiansheng Yang, Jian Wang, and Dongmei Fu. 2022. Dynamic Graph Reasoning for Multi-person 3D Pose Estimation. In Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547846

1 INTRODUCTION
The goal of multi-person 3D human pose estimation is to estimate the 3D coordinates of human joints from multiple human bodies in a monocular RGB image, which is a challenging and fundamental task. Recently, 3D human pose estimation has drawn a lot of attention because of its broad applications, such as human behavior understanding [22, 48], human-object interaction detection [20], and athletic training assistance [44]. Although remarkable progresses [24, 28, 37, 41, 43] have been achieved in 3D pose estimation, the challenges of depth ambiguity and occlusions remain.

Existing multi-person 3D pose estimation methods mainly include top-down, bottom-up, and single-stage strategies. Top-down strategy [28, 37, 38, 43] firstly predicts the human bounding box and the absolute depth of root points in each person and then conducts single-person 3D pose estimation in the bounding box. Bottom-up strategy [26, 27, 51, 54] firstly estimates the 3D coordinates for each human joint in an image and then assigns them to different human instances. Single-stage strategy [31, 55] predicts keypoints offsets for each place in the image to generate 3D poses. The top-down methods are more accurate but more costly since the human detection and the repeated stages of extracting features for each person. The bottom-up and single-stage approaches are more efficient but uncompetitive in accuracy. The occlusions, non-uniform scales, and variable depths of each person are more difficult to handle when the input is a whole image, which are our interests.

For 3D human pose estimation, the common way to decode 3D pose is based on heatmaps [23, 28, 41, 43, 54], which decodes 3D coordinates from 3D heatmaps by an isolated structure as Figure 1 (a). To mitigate occlusion problem, [23, 54] model context information in coordinate level and feature level, respectively. Recent works [31, 55] bridge these two levels to some extent by estimating the 3D keypoints offsets, directly decoding keypoints coordinates from root point with 3D offsets. For example, based on single root point, CenterNet [55] and SPM [31] estimate other 3D keypoints by star structure propagation as Figure 1 (b), which suffers the long-distance transmission problem. SPM further improves the information propagation path by tree structure decoding, as shown in Figure 1 (c). The tree structure brings accumulation error since the
The single-root decoding strategy, such as star and tree structure decoding, is unreliable since the depth ambiguity. Firstly, it is hard to calculate the localization of a new joint when its root joint is occluded. An incorrect starting point will result in a wrong whole 3D pose prediction. Secondly, the predicted depth from single root points is not robust since lacking the global context of the human body. Thirdly, the decoding graph is fixed for each person, which may fail to build the message propagation path when the middle nodes are occluded.

Based on the above analysis, we propose a dynamic graph reasoning (DGR) approach. As shown in Figure 1 (d), each joint in a person serves as a root keypoint to generate dense decoding paths. DGR reasons the dynamic decoding graph for each person from these dense decoding paths. Each message propagation path in the dynamic decoding graph is calculated by combining starting point, target point, and message path score, which is adaptively adjusted according to the occlusions in the input image. This mechanism enables the best decoding paths for each target joint. Although some works [23, 52, 53] use graph convolutional network or local human structure to learn better structural features, the learned body graph can’t be adjusted adaptively with the occlusion of the input image. But our DGR can self-adaptively adjust the decoding graph due to the multi-root decoding and dynamic graph reasoning mechanism.

To better predict the depth and build a reliable message propagation graph, we further propose a scale and depth aware refinement module. Inspired by the intuition of near big far small, which is a basic perspective principle, SADR concatenates multiple initial scale features and depth features and then generates refined scale and depth features. The refined scale-aware feature and depth-aware feature are beneficial to generating good root keypoints and building robust pose graphs.

Our main contributions can be summarized as follows:

- We argue that decoding 3D poses from a single root point is unreliable and propose a novel dynamic graph reasoning (DGR) method to decode multi-person 3D poses.
- We propose SADR to generate better root keypoints and reliable decoding graph representations by aggregating global depth and scale information.
- The proposed bottom-up approach outperforms even top-down methods and achieves new state-of-the-art results on three widely-used benchmarks: Human3.6M, MuPoTS-3D, and CMU Panoptic datasets.

2 RELATED WORK

2.1 2D Human Pose Estimation

Mainstream multi-person 2D pose estimation approaches include top-down and bottom-up methods. Top-down approaches [4, 39, 47] firstly conduct human detection. Then, they crop images and perform single-person human pose estimation for each human instance. Xiao et al. [47] propose a simple baseline for 2D pose estimation, which uses ResNet as the backbone and follows several up-sample layers to generate heatmaps. Sun et al. [39] propose HRNet to generate high-resolution representation. Bottom-up approaches [3, 17, 18] estimate the keypoints for all human instances in an image and then group them into multiple instances. Cao et al. [3] propose part affinity fields to group keypoints. Kocabas et al. [17] propose Multiposenet, a framework to finish detection, pose estimation, and grouping at the same time. The accuracy of bottom-up approaches is lower than top-down approaches since the different scales and low resolution of persons.

Some approaches [31, 42, 46, 55] discard the method of locating from heatmaps, but learn 2D offset to decode pose. Nie et al. [31] propose SPM to predict root points of the human body and offsets of each keypoints. The coordinates of keypoints can be obtained from the root points and offsets. DirectPose [42] and Point-set anchors [46] use deformable convolution to align the features of pose. These offset-based methods provide insight for 3D pose estimation.

2.2 3D Human Pose Estimation

For single person cases, the main-stream methods follow the architecture of top-down 2D pose estimation methods. They change the 2D pose regression heads to 3D heads, including *inference* and *lifting* methods. The *inference* methods [16, 25, 33, 40, 41] directly regress 3D pose from learned 3D heatmaps. For example, Sun et al. [41] conduct integral operation on 3D heatmaps to obtain 3D pose coordinates. The *lifting* methods [24, 29, 49, 52] firstly estimate

![Figure 1: (a) Heatmap-based approaches [28, 41] decode human poses by locating the point with maximum confidence on 3D heatmaps. Each node is isolated when decoding. (b) Star graph approaches [31, 55] start from single-root node, broadcasting information to other nodes by star graph, which suffers from long-distance information transmission problem. (c) Tree graph approach [31] starts from single-root node, hierarchically transferring the message from the parent node to the child node, which suffers from the occlusion problem and cumulative error. (d) Our dynamic graph reasoning (DGR) extracts dense decoding paths from multiple root keypoints, further reasons the best dynamic decoding graphs for each person in the image from the predicted dense decoding paths. The final 3D poses are decoded with the guidance of reasoned dynamic decoding graphs.](image)
2D pose by 2D pose estimator, and then lift 2D pose to 3D pose by a sample neural network, such as SRNet [52].

Recently, some works [28, 37, 38, 43, 51] study multi-person 3D pose estimation. Rogez et al. [37, 38] locate the human bounding box and generate a set of anchor poses for each human. The anchor poses are refined to the final pose by a regression module. Moon et al. [28] propose a top-down pipeline as multi-person 2D pose estimation. They locate the human bounding box and the depth of root keypoints in each bounding box and then conduct single-person 3D pose estimation in the bounding box. Zanfir et al. [51] propose MubyNet, which estimates keypoints and limb core at the same time and then integrates limb score to group keypoint into different persons. Wang et al. [43] propose hierarchical multi-person ordinal relations as an additional loss for depth learning. Cheng et al. [5] integrate top-down and bottom-up networks to relieve the problems of occlusion and close interactions.

2.3 Graph Reasoning in Pose Estimation

Since human joints can be naturally seemed as a graph structure, many works [2, 6, 23, 31, 35, 36, 53] try to utilize this information. Zhao et al. [53] take the human joints as the nodes of graph and then build a semantic graph convolutional network to learn joints relation. More works try to learn a better graph representation for human pose, such as spatial-temporal graph convolutional network [2], dynamic graph convolutional network [36], context pose [23] and so on. Other works [6, 52] study the local or global structure of the human body. However, these methods just learn graph representation for the human pose, the learned human graph can be regarded as scale map since the 2D offset indicates the scale of a person. \( \mathbf{M}_s \) is a 3D offset map, where each pixel records the 3D offsets from its 3D location to the corresponding \( K \) joints.

Based on those maps, our method outputs multi-person 3D pose by carrying out the following three parts successively, which are Scale and Depth Aware Refinement (SDAR), Multi-person Root Keypoints Decoding (MRKD), and Dynamic Graph Reasoning (DGR).

Firstly, SDAR refines the four maps mentioned above by integrating scale and depth information, and the refined maps are denoted as \( \mathbf{M}_h, \mathbf{M}_d, \mathbf{M}_s, \mathbf{M}_o \). Secondly, MRKD decodes multi-person root keypoints from those refined maps. \( \mathbf{M}_h, \mathbf{M}_d, \mathbf{M}_s \) are used to decode 3D root keypoints and \( \mathbf{M}_o \) is used to decode dense decoding paths. Finally, based on the root keypoints and predicted dense decoding paths, DGR reasons the dynamic decoding graphs for each person in the input image, to further decode robust multi-person 3D poses as equation 1 with the guidance of dynamic decoding graphs. The details of the three parts and the training process are introduced in the following.

3 APPROACH

Let \( p \) denotes the coordinate of a target joint, then the offset-based methods [31, 55] can be formulated as:

\[
p = p' + \Delta p
\]

where \( p' \) is the coordinate of a root joint, such as body center or the parent joint of \( p \) based on the human skeleton. \( \Delta p \) is a learned offset from the root joint to the target joint.

However, predicting 3D pose from single root is not robust since depth ambiguities and occlusions. Thus, we propose a new graph-based method, which takes full advantage of the context, scale, and depth information of multiple roots, and makes the 3D pose prediction more robust for handing depth ambiguity and occlusion.

3.1 Framework

The overview of our dynamic graph reasoning is illustrated in Figure 2. First of all, four maps and deep feature map \( F \) are extracted from the backbone network. These maps are denoted as \( \mathbf{M}_h, \mathbf{M}_d, \mathbf{M}_s, \mathbf{M}_o \) and \( \mathbf{M}_r \) of shapes \( (K+1, H, W), (K, H, W), (2, H, W) \) and \( (3K, H, W) \) respectively. \( K \) is the number of joint categories, \( H \) and \( W \) are the height and width of these maps. \( \mathbf{M}_h \) is the heat map for \( K \) joints and one body center (the midpoint of left hip and right hip). \( \mathbf{M}_d \) is the depth map for \( K \) joints. \( \mathbf{M}_s \) is a map where each pixel preserves the 2D offset from itself to the corresponding body center. And it can be regarded as scale map since the 2D offset indicates the scale of a person. \( \mathbf{M}_o \) is a 3D offset map, where each pixel records the 3D offsets from its 3D location to the corresponding \( K \) joints.

In the following, we denote \( \mathbf{M}_h, \mathbf{M}_d, \mathbf{M}_s, \mathbf{M}_o \) as \( \mathbf{M}_h, \mathbf{M}_d, \mathbf{M}_s, \mathbf{M}_o \). Firstly, SDAR refines the four maps mentioned above by integrating scale and depth information, and the refined maps are denoted as \( \mathbf{M}_h, \mathbf{M}_d, \mathbf{M}_s, \mathbf{M}_o \). Secondly, MRKD decodes multi-person root keypoints from those refined maps. \( \mathbf{M}_h, \mathbf{M}_d, \mathbf{M}_s \) are used to decode 3D root keypoints and \( \mathbf{M}_o \) is used to decode dense decoding paths. Finally, based on the root keypoints and predicted dense decoding paths, DGR reasons the dynamic decoding graphs for each person in the input image, to further decode robust multi-person 3D poses as equation 1 with the guidance of dynamic decoding graphs. The details of the three parts and the training process are introduced in the following.

3.2 Scale and Depth Aware Refinement

As demonstrated in Figure 3, SDAR firstly computes refined scale map and depth map by the following operations:

\[
\mathbf{M}_h = \text{Conv}_s(\text{Concat}(\mathbf{M}_h, \mathbf{M}_s)) \tag{2}
\]

\[
\mathbf{M}_d = \text{Conv}_d(\text{Concat}(\mathbf{M}_d, \mathbf{M}_s)) \tag{3}
\]

where \( \text{Conv}_s \) and \( \text{Conv}_d \) are sequential modules including convolution and normalization layers, and \( \text{Concat} \) is an operator for concatenating tensors in channel dimension. Equation 2 merges heat map into scale map, which makes \( \mathbf{M}_s \) can provide both scale information and attention cue for downstream operations. In Equation 3, depth map is refined to capture more global depth context from \( \mathbf{M}_o \).

Once got the refined scale map and depth map, we multiply them with the feature map \( F \) respectively, and then sum up these products as \( F_{sd} \). Based on this enhanced feature, the refined heat map and 3D offset map can be predicted by the following calculation:

\[
\mathbf{M}_h, \mathbf{M}_d = \text{Split}((\text{Conv}_s(F_{sd})))
\]

\[
F_{sd} = \mathbf{M}_s \odot F + \mathbf{M}_d \odot F \tag{4}
\]

where \( \text{Split} \) is also a sequential convolution module, \( \text{Split} \) is an operator for splitting tensor in channel dimension, and \( \odot \) means element-wise multiplication.

The final outputs are:

\[
\mathbf{M}_s = \mathbf{M}_s^R + \mathbf{M}_s^o \tag{5}
\]

where \( * \) is wildcard character for \( \{ h, s, d, o \} \).

In the following, \( \mathbf{M}_h, \mathbf{M}_t, \mathbf{M}_d, \mathbf{M}_o \) will serve for multi-person root keypoints decoding, while \( \mathbf{M}_h \) and \( \mathbf{M}_o \) will contribute for dynamic graph reasoning. After the refining by SDAR as Equation 4, the final multi-person pose predicted by DGR could be more precise since the heatmap and 3D offset map had been enhanced by scale and depth references.

3.3 Multi-person Root Keypoints Decoding

Before carrying out graph reasoning, we decode multi-person root 3D keypoints out of the maps outputted from SDAR. This process is
called multi-person root keypoints decoding (MRKD). At first, \(N \times K\) independent keypoints and \(N\) body centers can be detected from \(M_h\) by extracting local maxima, where \(N\) represents the number of persons and is obtained by finding \(N\) high-confidence points according to a confidence threshold. Meanwhile, the depth of each keypoint is the value of \(M_d\) at the corresponding 2D location. And then, we assign these keypoints to \(N\) persons. Let \(\{p_1, ..., p_m, ..., p_M\}\) as the \(M = N \times K\) predicted keypoints and \(\{c_1, ..., c_n, ..., c_N\}\) as the \(N\) predicted body centers from \(M_s\), where \(p_m\) and \(c_n\) are both 2D coordinates. Then each \(p_m\) is corresponding to a unique \(c_n\) by solving a distance matrix \(E^{M \times N}\) with hungarian algorithm, while the definition of each element in \(E\) is:

\[
E[m, n] = E(c_m, c_n) = |c_m|_{p_m + p_m}
\]  

(6)

where \(E\) represents calculating euclidean distance, and \(c_m\) is a regressed 2D body center coordinate from \(p_m\) depending on the semantic information at \(p_m\), while \(M_s|_{p_m}\) means the 2D offset from \(p_m\) to this center, which can be assigned as the 2D vector at point \(p_m\) on \(M_s\).

### 3.4 Dynamic Graph Reasoning

After obtaining root keypoints, the decoding graphs are inferred by conducting dynamic graph reasoning (DGR) for the final 3D pose estimation. Supposing that a person can be regarded as a undirected acyclic graph, denoted as \(G(\mathcal{P}, \mathcal{E})\), where \(\mathcal{P} = \{p^i, i \in [1, K]\}\) is the set of joints for this person, \(p^i\) is the 2D coordinate of the \(i^{th}\) joint. Here \(\mathcal{P}\) is initialized with the root joints detected from \(M_h\) and \(M_s\). While, \(\mathcal{E} = \{e^j, i \in [1, K], j \in [1, K]\}\) is the set of dense decoding paths and \(e^j\) means the decoding path from the \(i^{th}\) joint to the \(j^{th}\) joint. And it is valued by the 3D offset which is expressed as:

\[
e^j = M_o|_{p^i}^j
\]  

(7)

where \(M_o|_{p^i}^j\) means the 3D offset to the \(j^{th}\) target joint from \(M_o\) at the position of root joint \(p^i\).

The goal of DGR is to reason the best decoding paths \(\hat{\mathcal{E}}\) from dense decoding paths set \(\mathcal{E}\) to further construct decoding graph \(G(\mathcal{P}, \hat{\mathcal{E}})\) for each person in the input image. Intuitively, we can directly pick or drop candidate paths for generating \(\hat{\mathcal{E}}\). However, this hard selecting manner may not be optimal. Here, we adopt a soft manner that assigning a weight on each candidate path for reasoning the decoding graph. And the weights for all paths are inferred from heat map \(M_h\) and offset map \(M_o\). While the weighted decoding paths can be expressed as:

\[
\hat{\mathcal{E}} = \{(e^j, \mathcal{W}(p^i, p^j)), i \in [1, K], j \in [1, K]\}
\]  

(8)

where \(\mathcal{W}(p^i, p^j)\) means the corresponding path weight for \(e^j\), which can be calculated as:

\[
\mathcal{W}(p^i, p^j) = M_h|_{p^i}^j \cdot R(p^i, p^j)M_h|_{p^j}^{i^j}
\]  

(9)
where \( M_p^h \) and \( M_p^j \) serve as confidence scores which are the heat values of the \( i^{th} \) joint at point \( p_i \) on \( M_p \) and the \( j^{th} \) joint at point \( p_i \). While \( R(p_i, p_j) \) is a bone confidence formulated as:

\[
R(p_i, p_j) = \exp(-\frac{||M_p^{h_i}||_2 + \gamma(i, j))}{||M_p^{c_i}||_2})
\]

(10)

where \( h \) and \( c \) mean the joint index of head-top and mid-hip, respectively. \( M_p^{h_i} \) is the 3D offset from the head-top to mid-hip joint.

\[
\gamma(i, j) = ||\frac{M_p^{h_i}(i)}{\sigma(i, j)||_2 \text{ is a priori ratio, and } \sigma(i, j) \text{ is the average bone length between the } i^{th} \text{ and } j^{th} \text{ joint, counted from the training dataset, while } \sigma(h, c) \text{ is the average bone length between the } h^{th} \text{ joint and the } c^{th} \text{ joint. Thus, } R(p_i, p_j) \text{ indicates the confidence of predicted edge in the human structural graph, which appears as a production of instance-level propagation confidence and statistical priori propagation confidence.}

Given a pair of predicted root joint set \( \mathcal{P} \) and decoding graph \( G(\mathcal{P}, \mathcal{E}) \), the final 3D pose of a certain person can be decoded. Concretely, we firstly extend \( p_j \) in \( \mathcal{P} \) as \( \hat{p}_{3d} \) by concatenating \( p_j \) with its corresponding depth \( d_j \) predicted from \( M_d \). And the extended root joint set is denoted as \( \mathcal{P}_{3d} \). Then we decode \( G(\mathcal{P}_{3d}, \mathcal{E}) \) and update the 3D coordinates of each joint as:

\[
\hat{p}_{3d}^j = \frac{\sum_{i=1}^{K} W(p_i, p_j)(p_{3d}^i + \varepsilon^{ij})}{\sum_{i=1}^{K} W(p_i, p_j)}, \quad j = [1, K]
\]

(11)

where \( \hat{p}_{3d}^j \) represents the estimated 3D coordinates of the \( j^{th} \) target joint.

There are two reasons that the DGR can predict better 3D poses:

1) Compared with the single-root decoding mechanism, the depth value predicted by the multi-root decoding mechanism are more robust. 2) For each person, DGR builds a dynamic decoding graph, which can be adjusted by the path weights according to the deep feature of input image adaptively, even in the inference phase. As stated in Equation 9, the graph path confidence is determined by start point confidence, edge confidence, and endpoint confidence, which perceives the occlusion of the inputs image. This makes the 3D pose decoding more robust to occlusion conditions.

3.5 Loss Function

For training the network, we impose supervision both on the input and output maps of SDAR. The total loss is:

\[
L = L_k + \alpha L_s + \beta L_d + L_o,
\]

(12)

where \( \alpha = 0.1 \) and \( \beta = 0.1 \) are loss weights, and

\[
\begin{align*}
L_k &= L_{MSE}(M_p^h, \tilde{M}_h) + L_{MSE}(M_p^c, \tilde{M}_c) \\
L_s &= L_{L1}(M_p^h, \tilde{M}_h) + L_{L1}(M_p^c, \tilde{M}_c) \\
L_d &= L_{L1}(\delta(M_p^h), \tilde{M}_h) + L_{L1}(\delta(M_p^c), \tilde{M}_c) \\
L_o &= L_{L1}(M_{d}, \tilde{M}_d) + L_{L1}(M_o, \tilde{M}_o)
\end{align*}
\]

(13)

where \( \tilde{M}_i \) represents ground truth map. \( L_{MSE} \) is standard MSE loss. \( L_{L1} \) is L1 loss, and only pixels around body joints on \( \tilde{M}_c \), \( M_d \), and \( \tilde{M}_d \) are active for training. For training depth map, output transformation \( \delta(x) = 1/\text{sigmoid}(x) - 1 \) is applied on \( M_d \) and \( M_o \) before computing loss, following [8].

![Figure 3: The architecture of scale and depth refinement (SDAR). SDAR aggregates scale and depth information to learn scale-aware and depth-aware maps, further to refine keypoints heatmaps and 3D offsets. @ denotes concatenate operation, ⊕ denotes element-wise addition, ⊖ denotes element-wise multiplication.](image)

Table 1: The ablation study of SDAR and DGR on MuPoTS-3D dataset. The star and tree decoding approaches are introduced in [31]. The backbone of all results in this table is ResNet-34.

| Model          | Enhance Feature | Decoding Graph | 3DPCK | Δ   |
|----------------|-----------------|----------------|-------|-----|
| Baseline       | ×               | Star           | 75.4  | -   |
| Baseline       | ×               | Tree           | 76.7  | ↑1.7%|
| Ours(GR-M3D)   | ×               | DGR            | 77.9  | ↑3.3%|
| Ours(GR-M3D)   | SDAR            | DGR            | 78.6  | ↑4.2%|

4 EXPERIMENTS

4.1 Datasets and evaluation metrics

MuCo-3DHP and MuPoTS-3D MuCo-3DHP is a large-scale indoor multi-person training dataset [25], including 400K frames for training. For testing, MuPoTS-3D, which consists of real images with various camera poses, are collected from indoor and outdoor scenes. There are 20 real-world scenes in the MuPoTS-3D dataset, which are labeled with ground-truth 3D poses for multiple person subjects, making it a convincing benchmark to test the generalization ability of the 3D pose model.

The widely-used evaluation metric for multi-person 3D pose estimation is 3DPCK. If the Euclidean distance between predicted and ground-truth is smaller than the threshold (150mm), the prediction is marked as a correct prediction. \( PC_{Krel} \) measures relative pose accuracy after root alignment, and \( PC_{abs} \) measures absolute pose accuracy without root alignment. The area under the curve of 3DPCK over various thresholds is defined as AUC. To evaluate in the crowded scenes, we use Crowd Index to split hard cases and easy cases. \( 0 \leq \text{Crowd Index} \leq 1 \) is introduced in [19]. The bigger Crowd Index means the serious occlusion by human bodies.

Human3.6M Human3.6M [13] is the largest indoor benchmark for single-person 3D pose estimation, which consists of 3.6M video...
(a) Star structure decoding  
(b) Tree structure decoding  
(c) Dynamic Graph Reasoning  

Figure 4: The comparison of (a) star structure decoding, (b) tree structure decoding, and (c) our dynamic graph reasoning (DGR). DGR generates better results on these cases with occlusions or strange poses than star and tree structure decoding.

Table 2: The ablation study of our approach (GR-M3D) based on different backbone on MuPoTS-3D. HG2 denotes Hourglass with 2 blocks. RN denotes ResNet. HRN32 denotes HRNet-32. Baseline is with star structure.

| Model  | Backbone | Params(MB) | Time(s/img) | 3DPCK | Δ  |
|--------|----------|------------|-------------|-------|----|
| Baseline | RN34     | 32.1       | 0.022       | 75.4  | 4.2% |
| GR-M3D  |          | 34.4       | 0.026       | 78.6  |    |
| Baseline | RN50     | 40.6       | 0.029       | 77.1  | 3.1% |
| GR-M3D  |          | 42.9       | 0.034       | 79.5  |    |
| Baseline | RN101    | 59.6       | 0.041       | 79.1  | 2.8% |
| GR-M3D  |          | 61.9       | 0.045       | 81.3  |    |
| Baseline | HRN32    | 40.1       | 0.061       | 80.6  | 2.9% |
| GR-M3D  |          | 42.4       | 0.065       | 82.9  |    |
| Baseline | HG2      | 192.1      | 0.151       | 82.2  | 2.9% |
| GR-M3D  |          | 196.6      | 0.158       | 84.6  |    |

Table 3: The ablation study of GR-M3D on MuPoTS-3D dataset with different crowd indexes. Backbone is ResNet-50 here. Baseline is with star structure.

| Crowd Index | > 0.0 | > 0.3 | > 0.5 | > 0.7 |
|-------------|-------|-------|-------|-------|
| Baseline    | 77.1  | 76.4  | 74.8  | 73.2  |
| GR-M3D      | **79.5** | **78.8** | **77.9** | **77.5** |

4.3 Ablation study

Our baseline is the recently-developed offset-based method. They learn keypoint 3D offset and then decode human pose from single root keypoint by the corresponding star or tree pose graphs [31, 55].

Effectiveness of SDAR and DGR For the ablation study, we use ResNet-34 as the backbone. The ablation study of SDAR and DGR is shown in table 1. We can observe that decoding the 3D pose with the tree graph is better than the star graph, which gains a relative improvement of 1.7%. Our DGR achieves 77.9 3DPCK, which gains a relative improvement of 3.3%, compared with the star graph decoding method. We further use the scale and depth aware coordinates after root joint alignment and further rigid alignment (i.e., Procrustes analysis [11]).

CMU Panoptic CMU Panoptic [15] is a large-scale multi-person 3D human pose estimation dataset, captured by multiple cameras. It’s challenging to recover the multi-person 3D pose since heavy mutual occlusion. Following the settings of [50, 54], 9600 images of two cameras (16, 30) from four activities (Haggling, Mafia, Ultimatum, Pizza) as the testing set, and 160k images from different videos as training set. For fair comparison with [50, 54], MPJPE is the evaluation metric.

4.2 Implement details

GR-M3D is trained on 8 V100 GPUs with a batch size of 16/GPU and input resolution is 512 × 512. Adam optimizer is adopted and the initial learning rate is 5e-4, which decreases 10× at 60 and 90 epochs. The total epoch number is 110. Random flip, random occlusion, rotation, and color jittering are used, and the range of rotation is [−π, π]. The confidence threshold for root keypoints is 0.5. Unless otherwise specified, the backbone of GR-M3D is Hourglass network. Following previous works [9, 28, 54], additional 2D images in MPII [1] and COCO [21] are mixed into 3D datasets for training.
refinement module to improve the learned keypoints heatmaps, scale maps, depth maps, and 3D offset maps. SDAR achieves 78.6 3DPCK based on our DGR decoding approach. The whole gain of GR-M3D with SDAR and DGR is 3.2 3DPCK, a relative improvement of 4.2%, compared with the star structure decoding approach [31].

We also make a comparison of star, tree, and DGR decoding in Figure 4. As shown in Figure 4, the decoding methods based on star and tree structure fail in occlusion cases. Our DGR performs well on these challenging cases since the multi-root-based dynamic graph reasoning mechanism.

Different backbones To evaluate the effectiveness of SDAR and DGR on different backbones, we use several widely used networks (Hourglass [30], ResNet [12], HRNet-32 [39]) as backbone.

### Tables

#### Table 4: Comparison with state-of-the-art approaches on Human3.6M dataset.

| Method       | Direct. | Discuss | Eating | Greet | Phone | Photo | Pose | Purch. | Sitting | SittingD | Smoke | Wait | WalkD | WalkT | Avg. |
|--------------|---------|---------|--------|-------|-------|-------|------|--------|---------|----------|--------|------|-------|-------|------|
| Martinez et al. [24] | 50.5 | 53.7 | 53.9 | 46.8 | 50.0 | 61.9 | 83.2 | 82.5 | 80.0 | 94.4 | 87.6 | 86.7 | 61.5 | 63.4 | 59.1 |
| Sun et al. [41] | 47.5 | 47.7 | 50.2 | 43.8 | 46.4 | 58.9 | 65.7 | 49.4 | 55.8 | 47.8 | 38.9 | 49.0 | 43.8 | 49.6 |
| Moon et al. [28] | 50.5 | 55.7 | 50.1 | 51.7 | 53.9 | 46.8 | 50.0 | 61.9 | 68.0 | 52.5 | 49.9 | 41.8 | 56.1 | 46.9 | 53.3 |
| Cai et al. [2] | 46.5 | 48.8 | 47.6 | 50.9 | 52.9 | 61.3 | 48.3 | 45.8 | 59.2 | 64.4 | 51.2 | 48.4 | 53.5 | 39.2 | 41.2 |
| Zeng et al. [52] | 38.5 | 42.6 | 39.9 | 41.7 | 46.5 | 51.6 | 39.9 | 40.8 | 49.5 | 56.8 | 45.3 | 46.4 | 46.8 | 37.8 | 40.4 |
| Wehrbein et al. [45] | 36.3 | 42.8 | 39.5 | 40.0 | 43.9 | 48.8 | 36.7 | 44.0 | 51.0 | 63.1 | 44.3 | 40.6 | 44.4 | 34.9 | 46.7 |

**Ours**(GR-M3D) | 37.1 | 40.4 | 39.3 | 41.2 | 43.1 | 43.2 | 31.8 | 44.7 | 47.2 | 59.9 | 41.1 | 37.2 | 42.1 | 33.7 | 37.6 |

| Method       | PA MPJPE(mm) | Direct. | Discuss | Eating | Greet | Phone | Photo | Pose | Purch. | Sitting | SittingD | Smoke | Wait | WalkD | WalkT | Avg. |
|--------------|--------------|---------|---------|--------|-------|-------|-------|------|--------|---------|----------|--------|------|-------|-------|------|
| Martinez et al. [24] | 39.5 | 43.2 | 46.4 | 47.0 | 41.7 | 41.4 | 40.6 | 56.5 | 69.4 | 49.2 | 56.0 | 45.0 | 38.0 | 49.5 | 43.1 |
| Fang et al. [10] | 38.2 | 41.7 | 43.7 | 44.9 | 48.5 | 40.2 | 38.2 | 54.5 | 64.4 | 47.2 | 55.3 | 44.3 | 36.7 | 47.3 | 41.7 |
| Sun et al. [41] | 36.9 | 36.2 | 40.6 | 40.4 | 41.9 | 34.9 | 35.7 | 50.1 | 59.4 | 40.4 | 44.9 | 39.0 | 30.8 | 39.8 | 46.7 |
| Cai et al. [2] | 36.8 | 38.7 | 38.2 | 41.7 | 40.7 | 46.8 | 37.9 | 35.6 | 47.6 | 51.7 | 41.3 | 36.8 | 42.7 | 31.0 | 40.2 |
| Ma et al. [23] | 38.5 | 34.9 | 32.0 | 32.2 | 35.0 | 37.8 | 28.6 | 32.6 | 40.8 | 52.0 | 35.0 | 31.9 | 35.6 | 26.6 | 28.5 |
| Moon et al. [28] | 31.0 | 36.9 | 39.9 | 35.5 | 34.8 | 30.2 | 32.1 | 35.0 | 43.8 | 35.7 | 37.6 | 30.1 | 24.6 | 35.7 | 29.3 |
| Wehrbein et al. [45] | 27.9 | 31.4 | 29.7 | 30.2 | 34.9 | 37.1 | 27.3 | 28.2 | 39.0 | 46.1 | 34.2 | 32.3 | 33.6 | 26.1 | 27.5 |

**Ours**(GR-M3D) | 24.4 | 26.3 | 25.4 | 26.5 | 30.6 | 31.4 | 24.3 | 29.7 | 30.2 | 36.4 | 27.5 | 23.4 | 22.9 | 24.8 | 25.2 |

#### Table 5: Comparison with state-of-the-art approaches on MuPoTS-3D, a 3D multi-person pose estimation dataset. *"* shows that the results are not available. GR-M3D outperforms the SOTA methods. Bigger is better.

| Method       | Category | PCK\_rel | PCK\_abs | AUC\_rel |
|--------------|----------|----------|----------|----------|
| Lcr-net [37] |          | 62.4     | -        | -        |
| Lcr-net++ [38] | Top-down | 74.0     | -        | -        |
| HG-RCNN [7] |          | 74.2     | -        | -        |
| HMOR [43] |          | 82.0     | -        | -        |
| PoseNet [28] |          | 82.5     | 31.8     | 40.9     |
| ORPM [27] | Bottom-up | 69.8     | -        | -        |
| Xnect [26] | Bottom-up | 75.8     | -        | -        |
| SMAP [54] | Bottom-up | 80.5     | 38.7     | 42.7     |

**Ours**(GR-M3D) | Bottom-up | 84.6\*2.5% | 41.2\*6.5% | 44.1\*3.3% |

As shown in table 2, our GR-M3D outperforms baseline over a relative of 3% in different backbones. GR-M3D achieves 84.6 3DPCK on the MuPoTS-3D dataset, which is based on the Hourglass backbone. GR-M3D is lightweight and fast. GR-M3D only increases 2MB on the MuPoTS-3D dataset, which is based on the Hourglass backbone.

**Evaluation in crowded scenes** To evaluate the ability of GR-M3D to handle the occlusion cases, which are often seen in the crowded scenes, we evaluate GR-M3D on MuPoTS-3D with different Crowd Index. The bigger value of Crowd Index means more crowded cases. That represents the pose variances and occlusions are more serious. As shown in table 3, with the increase of Crowd Index, the 3DPCK of baseline decreases. At the same time, our GR-M3D outperforms baseline in all Crowd Index. Even in the
most challenging cases, with a crowd index of over 0.7, GR-M3D outperforms baseline with a relative improvement of 5.9%. These results show that our GR-M3D can handle the problems of pose variances and heavy occlusions.

4.4 Comparison with state-of-the-art methods
Following the settings of SOTA methods, we report the results of GR-M3D on the Human3.6M dataset. As shown in table 4, GR-M3D outperforms SOTA methods and achieves gains of 5% and 16% in MPJPE and PA MPJPE based on Hourglass backbone, respectively.

For multi-person cases, the commonly used dataset is the MuCo-3DHP and MuPoTS-3D dataset. We train GR-M3D on the MuCo-3DHP and test on the MuPoTS-3D. The results are shown in table 5. Based on Hourglass backbone, GR-M3D achieves the state-of-the-art results and obtains relative improvements of 2.5%, 6.5%, and 3.3% in $PCK_{rel}$, $PCK_{abs}$, and $AUC_{rel}$, respectively.

CMU Panoptic dataset is another widely-used benchmark for multi-person 3D pose estimation. Following [51, 54], we take experiments on this dataset and show results in table 6. Compared with state-of-the-art methods, GR-M3D obtains a relative improvement of 6.3% in MPJPE based on the Hourglass backbone.

4.5 Generalization in the wild
The images in MuPoTS-3D are collected in the wild scenes. Table 5 has shown that GR-M3D outperforms the state-of-the-art methods, which demonstrate the generalization ability of GR-M3D in the wild scenes. We also conduct experiments on COCO [21], a larger scale 2D pose estimation dataset. All of the images are collected in challenging, uncontrolled conditions. We directly predict the 3D pose on the COCO images since no 3D pose annotations. The model is based on Hourglass and trained on MuCo-3DHP. The results are shown in Figure 5. The visualization results in Figure 5 are from COCO, Human3.6M, MuPoTS-3D, and CMU Panoptic datasets, respectively. GR-M3D performs well on these challenging cases, even on the images collected in the outdoor scenes in the COCO dataset, which shows the generalization ability of GR-M3D.

5 CONCLUSION
We propose a novel bottom-up approach (GR-M3D) for 3D multi-person pose estimation, which mitigates occlusions and depth ambiguity by capturing more global context information. We firstly design a scale and depth refinement (SDAR) module to enhance the learned feature maps, to further generate better root keypoints and build robust message propagation paths. DGR reasons the dynamic decoding graphs from the predicted message propagation paths to decoding 3D poses. GR-M3D outperforms previous works and achieves state-of-the-art results on three widely-used benchmarks.

6 ACKNOWLEDGEMENT
This work was supported by the Scientific and Technological Innovation of Shunde Graduate School of University of Science and Technology Beijing (No. BK20AE004 and No.BK19CE017).
REFERENCES

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In CVPR. 3086–3093.

[2] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Jiansong Yuan, and Nadia Magnenat Thalmann. 2019. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In ICCV. 2272–2281.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In CVPR. 7291–7299.

[4] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2018. Cascaded pyramid network for multi-person pose estimation. In CVPR. 7103–7112.

[5] Yu Cheng, Bo Wang, Bo Yang, andobby T Tan. 2021. 3D Multi-Person Pose Estimation by Integrating Top-Down and Bottom-Up Networks. In CVPR. 7649–7659.

[6] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. 2019. Optimizing network structure for 3d human pose estimation. In ICCV. 2262–2271.

[7] Rashish Daharl, Nitesh B Gundavarapu, Rahul Mitra, Abhishek Sharma, Ganesh Ramakrishnan, and Arjun Jain. 2019. Multi-person 3d human pose estimation from monocular images. In 3DV. IEEE. 405–414.

[8] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In NeurIPS. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Eds.). https://proceedings.neurips.cc/paper/2014/file/7be7c9e171a690b5c6e5bae7a01-Paper.pdf

[9] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alfetta, and Rita Cucchiara. 2020. Compressed volumetric heatmaps for multi-person 3d pose estimation. In CVPR. 7204–7213.

[10] Hao-Shu Fang, Yuanlu Xu, Wenguang Wang, Xiaoabai Liu, and Song-Chun Zhu. 2018. Learning pose grammar to encode human body configuration for 3d pose estimation. In AAAI, Vol. 32.

[11] John C Gower. 1975. Generalized proximities matrix. Psychometrika 40, 1 (1975), 33–51.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In CVPR. 770–778.

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

[14] Ehsan Jalanghiri and Alan L Yuille. 2017. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In CVPR. 805–814.

[15] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. 2017. Panoptic studio: A massively interactive system for social interaction capture. TPAMI 41, 1 (2017), 190–204.

[16] Angjoo Kanarawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In CVPR. 7122–7131.

[17] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. 2018. Multiposenet: Fast multi-person pose estimation using pose residual network. In CVPR. 417–423.

[18] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. 2019. Pipaf: Composite fields for human pose estimation. In CVPR. 11977–11986.

[19] Jiefeng Li, Can Wang, Hao Zhu, Yubuan Mao, Hao-Shu Fang, and Cewu Lu. 2019. Crowds pose: Efficient crowds scene pose estimation and a new benchmark. In CVPR. 10863–10872.

[20] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. 2020. Detailed 2d-3d joint representation for human-object interaction. In CVPR. 10146–10157.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In CVPR. Springer, 740–755.

[22] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, and Wanli Ouyang. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In CVPR. 143–152.

[23] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Hai Ci, and Yiyou Zhou. 2021. Context Modeling in 3D Human Pose Estimation: A Unified Perspective. In CVPR. 6238–6247.

[24] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. 2017. A simple yet effective baseline for 3d human pose estimation. In ICCV. 2649–2649.

[25] Dushyant Mehta, Helge Rhodin, Dan Casal, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. 2017. Monocular 3d human pose estimation in the wild using improved cnn supervision. In 3DV. IEEE. 506–516.

[26] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. 2019. Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera. arXiv preprint arXiv:1907.00837 (2019).

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