

# COMPOSE: Hypergraph Cover Optimization for Multi-view 3D Human Pose Estimation

**Tony Danjun Wang**

*School of Computation, Information, and Technology, Technical University of Munich, Germany*

*tony.wang@tum.de*

**Tolga Birdal**

*Department of Computing, Imperial College London, United Kingdom*

*t.birdal@imperial.ac.uk*

**Nassir Navab**

*School of Computation, Information, and Technology, Technical University of Munich, Germany  
Munich Center for Machine Learning, Germany*

*nassir.navab@tum.de*

**Lennart Bastian**

*School of Computation, Information, and Technology, Technical University of Munich, Germany  
Munich Center for Machine Learning, Germany*

*lennart.bastian@tum.de*

## Abstract

3D pose estimation from sparse multi-views is a critical task for numerous applications, including action recognition, sports analysis, and human-robot interaction. Optimization-based methods typically follow a two-stage pipeline, first detecting 2D keypoints in each view and then associating these detections across views to triangulate the 3D pose. Existing methods rely on mere pairwise associations to model this correspondence problem, treating global consistency between views (i.e., cycle consistency) as a soft constraint. Yet, reconciling these constraints for multiple views becomes brittle when spurious associations propagate errors. We thus propose COMPOSE, a novel framework that formulates multi-view pose correspondence matching as a hypergraph partitioning problem rather than through pairwise association. While the complexity of the resulting integer linear program grows exponentially in theory, we introduce an efficient geometric pruning strategy to substantially reduce the search space. COMPOSE achieves improvements of up to 23% in average precision over previous optimization-based methods and up to 11% over self-supervised end-to-end learned methods, offering a promising solution to a widely studied problem.

## 1 Introduction

Human pose estimation is a fundamental task in computer vision, yet its deployment in safety-critical scenarios remains a significant challenge. While the research community has primarily focused on monocular settings due to the availability of large-scale annotated datasets, these methods inherently lack depth information and suffer from severe occlusions [ZWC<sup>+</sup>23]. This limitation is unacceptable in high-stakes real-world applications such as collaborative human-robot interaction [GS08] and operating room monitoring [WBC<sup>+</sup>25], where precise spatial awareness is required to ensure safety. To achieve the necessary robustness and geometric accuracy, standard practice involves deploying calibrated multi-camera setups. By capturing activities from diverse viewpoints, these setups allow for geometric triangulation, thereby resolving the depth ambiguities inherent to single-view imaging.

Multi-view 3D human pose estimation has historically relied on these geometric multi-view consistency constraints [NOT25]. Seminal approaches employ optimization to lift 2D detections from multiple monocular views into 3D space via triangulation [BAA<sup>+</sup>14]. However, purely geometric methods can be highly sensitive to 2D detection ambiguities and noisy keypoints, leading to suboptimal performance in complex, crowded scenes.

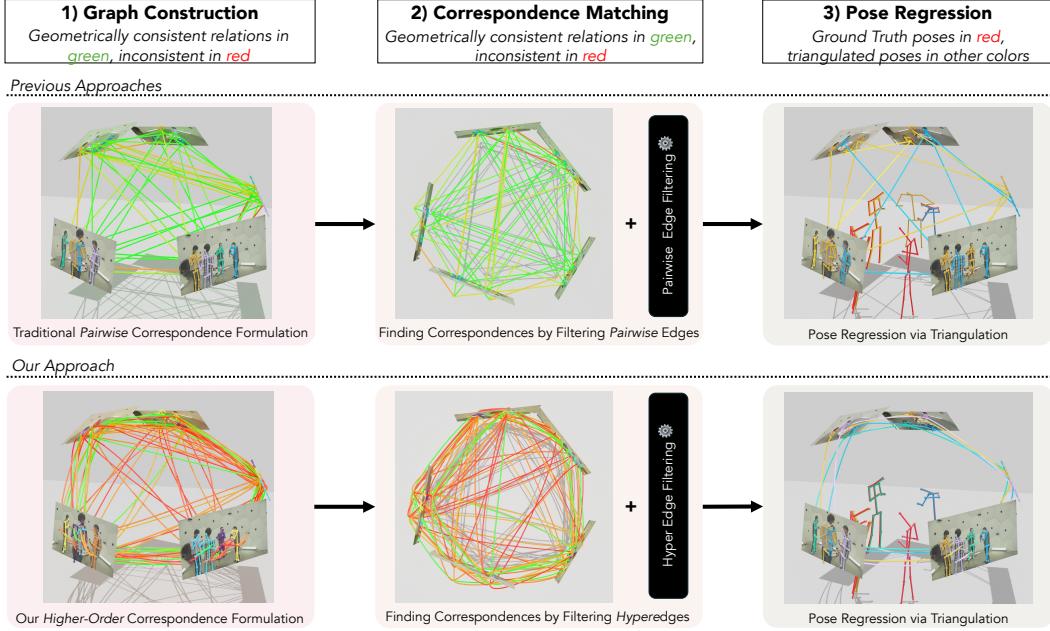


Figure 1: **Top:** Traditional approaches rely on pairwise geometric constraints [DFJ<sup>+</sup>22, ZAY<sup>+</sup>20, WJL<sup>+</sup>21]. As illustrated, these methods generate pairwise associations that, while *locally* consistent between two views, often fail to form a *globally* coherent structure. Consequently, algorithms face the difficult task of reconciling these locally plausible but globally conflicting edges to recover the correct 3D poses. **Bottom:** We propose a hypergraph formulation that jointly models higher-order relationships across views. We re-frame correspondence matching as a *hypergraph partitioning* problem, where hyperedges encode multi-view consistency. This global formulation effectively resolves ambiguities by enforcing consensus across the entire set of views.

The advent of deep learning revolutionized 2D pose estimation, providing robust keypoint detections that serve as improved inputs for geometric lifting. Despite these advances in 2D backbones, the community has largely shifted toward end-to-end learning-based 3D frameworks [TWZ20, YZW<sup>+</sup>22, LZW<sup>+</sup>24]. While these approaches yield impressive results, they require large annotated datasets. However, obtaining ground-truth 3D annotations is labor-intensive and technically challenging; since skeletal joints are internal to the human body, they cannot be directly observed or accurately annotated from surface scans alone. Furthermore, learning-based models frequently suffer from domain gaps, failing to generalize to unseen environments or novel camera configurations [LZW<sup>+</sup>24, CGD25]. Methods that can generalize without reliance on precise 3D supervision are thus highly sought after.

In this paper, we propose COMPOSE, a novel modeling framework for multi-view 3D human pose estimation that effectively captures the complex relations between 2D observations and 3D human poses (see Figure 1). Previous approaches tend to model multi-view relationships by *synchronizing* many dyadic (pairwise) relations, attempting to recover a cycle-consistent matching between views [DFJ<sup>+</sup>22, HWB<sup>+</sup>21, HWB<sup>+</sup>21, BS19, CSMP25]. However, such methods are inhibited by their reliance on operating between pairwise matches computed in isolation at a time (see top of Figure 1). This makes it challenging to resolve ambiguities when individual views are occluded or noisy [HWY<sup>+</sup>22].

To alleviate these challenges, we introduce a hypergraph-based formulation that extends beyond pairwise relationships to capture higher-order interactions among multiple views simultaneously. By treating consistent sets of 2D detections across views as hyperedges, our model enforces a holistic consensus across all cameras simultaneously, making it significantly more robust to outliers in individual frames (see Figure 1, where hyperedges better express global geometric consistency). With a purely optimization-based approach, our method successfully regresses accurate 3D human poses from multi-view 2D observations, even in the presence of severe occlusions. Furthermore, our experiments demonstrate that our hypergraph-based approach outperforms state-of-the-art optimization-based and self-supervised learning methods.

The main contributions of this paper are summarized as follows:

- We propose COMPOSE, a novel hypergraph formulation for multi-view 3D human pose estimation that advances the paradigm from pairwise association between camera views to considering detections from all cameras simultaneously; this is cast as a partitioning problem over a higher-order graph.
- We introduce a robust optimization objective and a geometrically guided pruning strategy that enables the efficient solution of the theoretically exponential correspondence problem via Integer Linear Programming (ILP).
- We demonstrate through extensive experiments that our optimization-based method outperforms state-of-the-art optimization-based as well as self-supervised learning-based methods on several standard benchmarks.

## 2 Related Works

### 2.1 Optimization-Based Approaches

3D multi-view multi-human pose estimation was first established through optimization-based approaches that lift 2D priors (e.g., off-the-shelf 2D human pose estimator) into 3D by triangulating 2D correspondences across views. As a pioneering work, Belagiannis et al. [BAA<sup>+</sup>14] introduce a conditional random field framework (3D Pictorial Structures) that uses a discrete state space and multi-view potential functions to resolve identity and body part ambiguities across multiple camera views. Dong et al. [DFJ<sup>+</sup>22] establish multi-view correspondences by formulating the multi-way matching problem as a convex optimization problem that simultaneously clusters 2D detections across all views using appearance similarity and geometric consistency cues while enforcing a cycle-consistency constraint. To address occlusions and crowded scenes, Zhou et al. [ZSW<sup>+</sup>22] introduce multi-view association at the level of partial skeleton proposals instead of body-level. Zhang et al. [ZAY<sup>+</sup>20] introduce the temporal dimension to the task and propose a spatio-temporal graph formulation for both spatial and temporal associations. These optimization-based methods are generally efficient and require minimal computational resources. However, compared to learnable methods, they still struggle to handle noisy 2D detections and occlusions.

### 2.2 Learning-Based Approaches

Currently, the prevalent body of work in 3D multi-view multi-human pose estimation focuses on approaches that learn to regress 3D human poses. Early learning-based methods encode the environment as 3D voxel grids and use 3D CNNs to directly regress 3D human poses from the voxelized representation [IBLM19, TWZ20, CT25, SCP24]. However, these methods are generally computationally expensive due to the cubic complexity of 3D CNNs. Moreover, voxelized representations are prone to overfitting to specific camera setups and do not generalize well to unseen camera configurations [LZW<sup>+</sup>24]. To address these issues, recent methods adopt a more efficient and flexible approach by directly inferring 3D human poses by projecting 3D hypotheses onto 2D image planes and leveraging 2D image features [WZC<sup>+</sup>21, LZW<sup>+</sup>24, CGD25]. These methods have shown better generalization to unseen camera configurations while being more computationally efficient.

Despite the success of learning-based methods, they are still fundamentally dependent on large-scale, 3D-annotated datasets, which are scarce and laborious to obtain. Therefore, optimization-based approaches remain a critical alternative, offering robust generalization in diverse settings without requiring expensive 3D supervision.

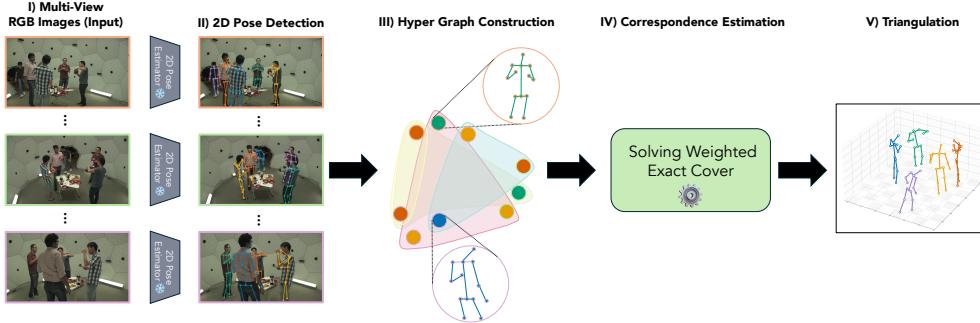


Figure 2: An overview of COMPOSE. Multi-view images are taken as input (I). First, we employ an off-the-shelf 2D pose estimator to extract 2D keypoints (II). Next, we construct our weighted hypergraph (III). We then solve the weighted exact cover problem to optimally partition the graph and establish unique correspondences (IV). Finally, we triangulate the 3D keypoints from these correspondences to obtain our final 3D human poses (V).

### 3 Methodology

COMPOSE tackles 3D multi-view multi-person pose estimation by first detecting 2D poses in each view independently using off-the-shelf detectors, as is standard in the literature [BAA<sup>+</sup>14, PGS19, DFJ<sup>+</sup>22], and subsequently establish inter-view correspondences to triangulate the 2D detections into 3D poses via a novel optimization formulation over a higher-order graph. Figure 2 illustrates the pipeline of our proposed framework. We now present our formulation, formally.

#### 3.1 Problem Setting

We aim to recover 3D multi-person poses from a sparse set of RGB cameras (datasets for this task typically contain 3–8 views; see [NOT25, Table 2]). Formally, we seek to estimate the set of 3D human poses  $\mathcal{P} = \{P_k\}_{k=1}^K$  for  $K$  individuals. Each pose  $P_k \in \mathbb{R}^{J \times 3}$  comprises  $J$  body joints.

Let  $\mathcal{U}_v = \{U_i^v\}$  denote the set of detected 2D poses in view  $v$ , where  $U_i^v$  represents the  $i$ -th candidate pose. Accordingly, the 2D pixel coordinate of joint  $j$  for this candidate is given by  $\mathbf{u}_{i,j}^v \in \mathbb{R}^2$ . The set of all detected poses is denoted as  $\mathcal{U} = \bigcup_v \mathcal{U}_v$ .

To estimate  $\mathcal{P}$ , we must partition the set  $\mathcal{U}$  into disjoint subsets, where each subset corresponds to a unique underlying individual observed across views. This partitioning allows us to triangulate the 3D joint positions for each person. We frame this as a global optimization problem, where we seek to identify the optimal partition that maximizes geometric consistency across all views simultaneously.

#### 3.2 A Higher-Order Formulation

**Nodes (2D human joints).** We first extract the 2D human poses as described in the previous section. The extracted joints serve as the nodes of our hypergraph. We employ VIT-pose [XZZT22], a standard top-down 2D pose detector, to process each view independently, yielding the set of poses  $\mathcal{U}_v$  for each view  $v$ . These form the vertex set  $\mathcal{V} = \mathcal{U}$  of our hypergraph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ .

**Correspondence Construction.** Each hyperedge  $e \in \mathcal{E}$  represents a *hypothesis* for one unique 3D person, linking a set of 2D observations across multiple views. Theoretically, the set of all possible hyperedges  $\mathcal{E}_{all}$  includes every combination of detections across all subsets of views, which grows exponentially in  $V$ .

**Proposition 1.** *The total number of potential hyperedges  $M$  is given by:*

$$M = \sum_{\emptyset \neq S \subseteq \{1, \dots, V\}} \prod_{v \in S} n_v = \prod_{v=1}^V (1 + n_v) - 1$$

where  $S$  represents a subset of views and  $n_v = |\mathcal{U}_v|$  is the number of detected poses in view  $v$ . This implies exponential growth with respect to the number of views  $V$ , yielding a complexity  $\mathcal{O}((N+1)^V)$  where  $N$  is the total number of individuals present.

*Proof.* Assuming  $n_v = N$  for all views, the total number of hyperedges is:

$$M = \sum_{k=1}^V \binom{V}{k} N^k = \sum_{k=0}^V \binom{V}{k} N^k - 1 = (1+N)^V - 1 \in \mathcal{O}((N+1)^V)$$

where the second equality follows from the binomial theorem.  $\square$

Prop. 1 makes a naive optimization over  $\mathcal{E}_{all}$  intractable for large  $V$ . While multi-view camera estimation methods are typically limited to a modest number of views [NOT25], we observe that many hyperedges are geometrically implausible. However, by looking at geometric consistency measures for a given hyperedge, it is possible to gauge its validity and potentially prune the search space accordingly. Next, we present our optimization, which leverages such consistency measures.

### 3.3 Recovering Multi-view 3D Human Poses

We formulate the correspondence problem as a *Weighted Exact Cover* over the hypergraph [Kar09]. Our goal is to select the smallest subset of hyperedges that explains every observed node exactly once, while maximizing the total confidence.

Let  $x_e$  be a binary decision variable for each hyperedge  $e \in \mathcal{E}$ :

$$x_e = \begin{cases} 1, & \text{if hyperedge } e \text{ is selected,} \\ 0, & \text{otherwise.} \end{cases}$$

Then the optimization objective can be formulated as:

$$\begin{aligned} \max_{\{x_e\}} \quad & \sum_{e \in \mathcal{E}} (s(e) - \gamma)x_e \\ \text{s.t.} \quad & \sum_{e: u \in e} x_e = 1, \quad \forall u \in \mathcal{V}, \\ & x_e \in \{0, 1\}, \quad \forall e \in \mathcal{E}. \end{aligned}$$

Here,  $\gamma > 0$  acts as a sparsity regularizer,  $\gamma$  penalizes the number of edges in the recovered partition, encouraging the solver to prefer larger hyperedges that explain consistency across more views, and  $s(e)$  is a confidence score associated with each hyperedge, to be specified in what follows.

**Geometrically Consistent Optimization.** To gauge the plausibility of a hyperedge candidate, we define a geometric consistency cost  $\mathcal{C}(e)$ . Notably, our graph formulation is agnostic to the specific choice of this metric, thus allowing for flexibility depending on the available calibration data. For instance, in weakly- or uncalibrated settings [LXK<sup>+</sup>24, HSZ<sup>+</sup>21],  $\mathcal{C}(e)$  could be formulated using the Sampson error derived from the Fundamental matrix or via Trifocal tensors for triplet constraints [HZ03].

In this vein of literature, it is commonly assumed that cameras are calibrated and co-registered [NOT25]. We thus employ the reprojection error as our primary instantiation of  $\mathcal{C}(e)$ . Formally, the optimal reprojection error for a set of correspondences within a hyperedge  $e$  is defined as:

$$E_{\text{geom}}^* = \min_{\{\hat{\mathbf{y}}_j\}} \frac{1}{J} \sum_{j=1}^J \sum_{U_i^v \in e} \|\pi(K_v, R_v, t_v, \hat{\mathbf{y}}_j) - \mathbf{u}_{i,j}^v\|^2$$

where  $\hat{\mathbf{y}}_j \in \mathbb{R}^3$  is the triangulated 3D position of joint  $j$ . We approximate this efficiently using closed-form algebraic triangulation ( $\hat{\mathbf{y}}_{\text{DLT},j}$ ):

$$\mathcal{C}(e) \approx \frac{1}{J} \sum_{j=1}^J \sum_{U_i^v \in e} \|\pi(K_v, R_v, t_v, \hat{\mathbf{y}}_{\text{DLT},j}) - \mathbf{u}_{i,j}^v\|^2$$

Here, the cost  $\mathcal{C}(e)$  is averaged over all  $J$  body joints. This cost is finally converted into a confidence score  $s(e)$  using an exponential kernel, where  $\lambda$  controls sensitivity:

$$s(e) = \exp(-\lambda \cdot \mathcal{C}(e))$$

**Graph Construction and Pruning.** Using our metric, we construct a tractable set of hyperedges  $\mathcal{E} \subset \mathcal{E}_{all}$  via pruning. Instead of instantiating all combinations, we discard any hyperedge where the geometric cost exceeds a predefined threshold  $\tau$  (i.e.,  $\mathcal{C}(e) > \tau$ ). This pruning strategy focuses the optimization on the sparse set of physically plausible hypotheses. Additionally, to account for single-view detections (e.g., heavily occluded individuals), we include singleton hyperedges with a fixed prior score  $s(e) = s_{single}$ .

**Optimization Strategy.** As our formulation corresponds to the *weighted exact cover* problem, it is NP-complete [Kar09]. We solve it using Integer Linear Programming (ILP) with a branch-and-bound algorithm [LD60] via the PuLP library [MOD11]. Although the worst-case complexity is exponential, our geometric pruning strategy significantly reduces the search space, leading to efficient convergence in practice (see subsection 4.5).

## 4 Experiments and Results

### 4.1 Datasets

We evaluate our method on three widely used public datasets that provide synchronized multi-view RGB images and 3D human pose ground truth in global coordinates.

**CMU Panoptic [JLT<sup>+</sup>15]** is a large-scale dataset captured in an indoor studio environment with a massive multi-camera system. Following related work, we adopt the standard train/test split protocol and camera setup (cameras 3, 6, 12, 13, and 23) established in prior studies for our task [TWZ20, CKJ23, LZW<sup>+</sup>24, YZW<sup>+</sup>22].

**Shelf Dataset [BAA<sup>+</sup>14]** captures a small indoor environment with extensive occlusions, where four people interact with a shelf. It consists of synchronized footage from five calibrated cameras. We use the standard evaluation protocol, testing on the provided testing frames and performing hyperparameter tuning on the remaining frames.

**Campus Dataset [BAA<sup>+</sup>14]** was captured in an outdoor environment using three calibrated cameras and features multiple people walking in an open courtyard. It evaluates the method’s robustness under less controlled lighting conditions and with fewer camera views.

### 4.2 Experiments

We compare COMPOSE against both learning-based and optimization-based baselines. We use standard evaluation metrics: Average Precision (AP) at various thresholds, Recall at 500mm, and Mean Per Joint Position Error (MPJPE).

For MvPose [DFJ<sup>+</sup>22], the leading optimization-based baseline, we utilize the official implementation and evaluate their method using the same 2D keypoints as our method to ensure a fair comparison. For the remaining baselines, we report the results as published in their respective publications.

### 4.3 Quantitative Results

**CMU Panoptic.** Table 1 presents the quantitative comparison on the CMU Panoptic dataset [JLT<sup>+</sup>15]. COMPOSE outperforms competing optimization-based approaches and surpasses the self-supervised learning-based methods in several metrics. Against fully-supervised methods, our approach remains highly competitive in less strict metrics.

Specifically, compared to the strongest optimization-based baseline, MvPose [DFJ<sup>+</sup>22], COMPOSE achieves higher accuracy across all metrics. We observe a substantial improvement in Average

Table 1: Quantitative comparison on the CMU Panoptic dataset [JLT<sup>+</sup>15]. We report Average Precision (AP) in mm, Recall, and Mean Per Joint Position Error (MPJPE) in mm. <sup>†</sup> uses 9 temporal frames as input. <sup>‡</sup> uses the same 2D keypoint detector as our method. Best results per supervision category (full-, self-, and optimization-based) are highlighted in **blue**, **orange**, and **green**.

Method	Average Precision (AP) ( $\uparrow$ )				Recall ( $\uparrow$ ) @500	Error ( $\downarrow$ ) MPJPE
	25	50	100	150		
<i>Fully-Supervised</i>						
VoxelPose [TWZ20]	83.59	98.33	99.76	99.91	–	17.68
Plane Sweep Pose [LL21]	92.12	98.96	<b>99.81</b>	99.84	–	16.75
MvP [WZC <sup>+</sup> 21]	92.28	96.60	97.45	97.69	–	15.76
Faster VoxelPose [YZW <sup>+</sup> 22]	85.22	98.08	99.32	99.48	–	18.26
Wu <i>et al.</i> [WJL <sup>+</sup> 21]	93.93	98.93	99.78	99.90	<b>99.97</b>	15.63
TEMPO [CKJ23]	89.01	<b>99.08</b>	99.76	<b>99.93</b>	–	14.68
MVGFormer [LZW <sup>+</sup> 24]	92.32	97.93	99.32	99.55	99.86	15.99
VoxelPose + 3DSA [CT25]	<b>94.20</b>	98.49	99.21	99.31	–	<b>13.98</b>
MV-SSM [CGD25]	93.50	–	–	–	–	15.70
<i>Self-Supervised</i>						
SelfPose3d [SCP24]	55.13	<b>96.44</b>	<b>98.46</b>	<b>98.98</b>	<b>99.60</b>	24.47
DSP <sup>†</sup> [LZ25]	<b>57.60</b>	86.10	94.00	–	–	<b>23.10</b>
<i>Optimization-Based</i>						
ACTOR [PGS19]	–	–	–	–	–	168.40
MvPose <sup>‡</sup> [DFJ <sup>+</sup> 22]	37.63	95.70	97.84	98.28	99.60	26.46
COMPOSE (Ours)	<b>54.66</b>	<b>97.27</b>	<b>98.94</b>	<b>99.17</b>	<b>99.83</b>	<b>23.62</b>

Table 2: Quantitative comparison on the Shelf [BAA<sup>+</sup>14] and Campus [BAA<sup>+</sup>14] datasets (PCP %). <sup>‡</sup> reproduced using the same 2D keypoint detector as our method. Best results per supervision category (full-, self-, and optimization-based) are highlighted in **blue**, **orange**, and **green**.

Method	Shelf (PCP %) ( $\uparrow$ )				Campus (PCP %) ( $\uparrow$ )			
	Actor 1	Actor 2	Actor 3	Avg.	Actor 1	Actor 2	Actor 3	Avg.
<i>Fully Supervised</i>								
Ershadi <i>et al.</i> [ENKS18]	93.3	75.9	94.8	88.0	94.2	92.9	84.6	90.6
VoxelPose [TWZ20]	99.3	94.1	97.6	97.0	97.6	93.8	<b>98.8</b>	96.7
Wu <i>et al.</i> [WJL <sup>+</sup> 21]	99.3	<b>96.5</b>	97.3	<b>97.7</b>	–	–	–	–
MvP [WZC <sup>+</sup> 21]	99.3	95.1	<b>97.8</b>	97.4	<b>98.2</b>	94.1	97.4	96.6
Faster VoxelPose [YZW <sup>+</sup> 22]	<b>99.4</b>	96.0	97.5	97.6	96.5	94.1	97.9	96.2
TEMPO [CKJ23]	99.3	95.1	<b>97.8</b>	97.4	97.7	<b>95.5</b>	97.9	<b>97.3</b>
<i>Self-Supervised</i>								
SelfPose3d	<b>97.2</b>	<b>90.3</b>	<b>97.9</b>	<b>95.1</b>	<b>92.5</b>	<b>82.2</b>	<b>89.2</b>	<b>87.9</b>
<i>Optimization-Based</i>								
3DPS [BAA <sup>+</sup> 14]	75.3	69.7	87.6	77.5	93.5	75.7	84.4	84.5
MvPose [DFJ <sup>+</sup> 22]	98.8	<b>94.1</b>	<b>97.8</b>	<b>96.9</b>	97.6	93.3	98.0	96.3
MvPose <sup>‡</sup> [DFJ <sup>+</sup> 22]	99.5	91.6	96.3	95.8	98.4	93.4	95.7	95.8
COMPOSE (Ours)	<b>99.8</b>	92.4	96.3	96.2	<b>99.4</b>	<b>94.3</b>	<b>98.1</b>	<b>97.3</b>

Precision ( $AP_{25}$ ), increasing from 37.63 to 54.66. Furthermore, we achieve a lower MPJPE of 23.62mm compared to 26.46mm for MvPose, and a higher recall of 99.83% compared to 99.60%.

While fully-supervised methods like VoxelPose [TWZ20] and its variants generally achieve lower error rates due to the available 3D ground truth supervision, our method bridges the gap without requiring any 3D annotations. Notably, COMPOSE outperforms recent self-supervised approaches like SelfPose3d [SCP24] in almost all metrics. Compared to DSP [LZ25], we achieve competitive results; while DSP reports a slightly lower error (23.10mm vs. 23.62mm), our method maintains comparable precision and recall without heavy training requirements.

**Shelf and Campus.** To evaluate the generalization capability of COMPOSE, we report the

Percentage of Correct Parts (PCP) on the Shelf and Campus datasets in Table 2.

On the **Shelf dataset**, COMPOSE achieves an average PCP of 96.2%, which is comparable to state-of-the-art fully-supervised methods and outperforms the self-supervised baseline SelfPose3d (95.1%). When compared to the optimization-based baseline MvPose (96.9%), our method remains competitive, particularly on Actor 1, where we achieve 99.8% accuracy. We also note that the Shelf dataset contains occasional inaccuracies in the ground truth annotations. In subsection 4.4, we provide qualitative examples where our method’s predictions align better with the image evidence than the provided ground truth.

On the **Campus dataset**, COMPOSE demonstrates strong robustness, achieving the highest average PCP of 97.3% among optimization-based methods, tying with the fully-supervised method TEMPO [CKJ23]. Notably, we outperform MvPose (96.3%) and significantly surpass the self-supervised SelfPose3d (87.9%).

#### 4.4 Qualitative Results

We present qualitative visualizations to further validate the effectiveness and robustness of COMPOSE.

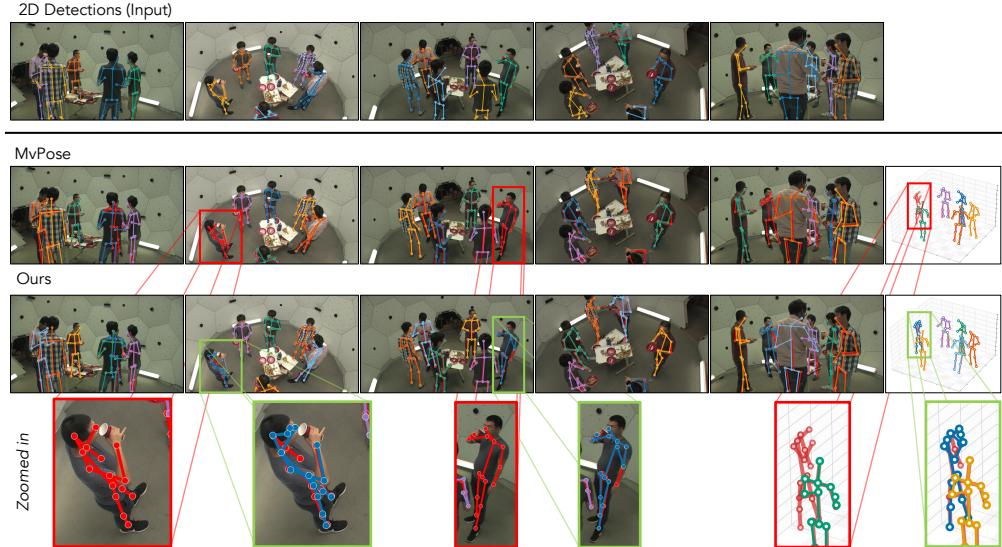


Figure 3: Qualitative results on the CMU panoptic dataset [JLT<sup>+</sup>15]. We show input 2D detections obtained from ViTPose [XZZT22] and the 3D predictions from MvPose [DFJ<sup>+</sup>22] and our method. 3D ground truths are visualized in red, predictions are visualized in other colors. As highlighted, MvPose fails to predict the correspondences for one person, while our method successfully reconstructs all individuals.

**Comparison on CMU Panoptic.** Figure 3 compares the reconstruction results of COMPOSE against MvPose [DFJ<sup>+</sup>22] on the CMU Panoptic dataset. We utilize 2D detections from ViTPose [XZZT22] as input for both methods. As illustrated in the zoomed-in regions, MvPose fails to establish correct correspondences for the highlighted individual, resulting in a missing reconstruction. In contrast, our method successfully disentangles the multi-view information and accurately reconstructs all individuals in the scene.

**Robustness on Shelf and Ground Truth Inaccuracies.** We further evaluate COMPOSE on the Shelf dataset [BAA<sup>+</sup>14]. Figure 4 visualizes our predictions alongside the publicly provided ground truth annotations. Evidently, the Shelf dataset contains instances of unreliable ground truth. As shown in the highlighted examples in Figure 4, the ground truth annotations (dashed red lines) significantly deviate from the actual position of the actors visible in the image. Our method (solid lines), however, produces poses that are visually consistent with the image evidence. This discrepancy penalizes the

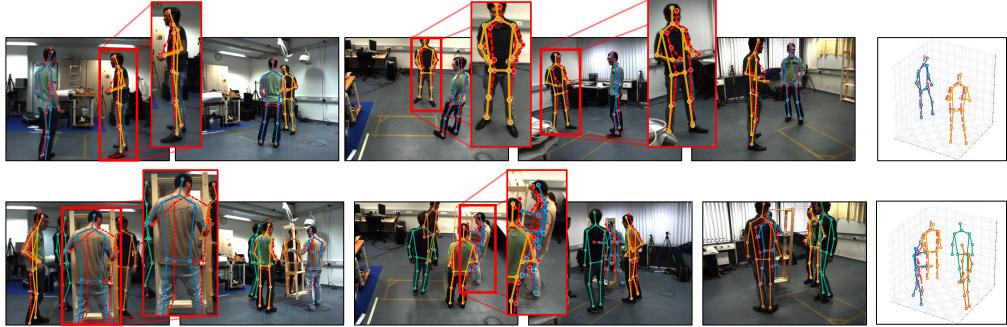


Figure 4: Qualitative results on the Shelf dataset [BAA<sup>+</sup>14] (frame 305 and 544). We visualize the ground truth (dashed red lines) and our predictions (solid colored lines). Note that for the highlighted actors, the ground truth annotations deviate significantly from the image evidence, while our method recovers the visually correct poses. Best viewed in color and zoomed in.

quantitative evaluation; for instance, in the top example, despite the visual accuracy, our method receives a PCP of 0% for the right lower and upper arms due to the incorrect ground truth. Similarly, for the bottom example, the prediction results in a PCP of 0% for the right lower arm, upper arm, and head.

## 4.5 Run Time Analysis

Table 3: ILP solving runtime analysis on CMU Panoptic [JLT<sup>+</sup>15] (standard 5 camera setup) with respect to the number of people in the scene. We report the mean runtime in ms, with the standard error.

# People	1	2	3	4	5	6	Avg.
Time (ms)	$9.73 \pm 0.27$	$14.30 \pm 0.13$	$21.09 \pm 0.24$	$30.07 \pm 2.45$	$29.33 \pm 0.28$	$34.56 \pm 0.15$	$22.70 \pm 0.20$

In Table 3, we report the run time performance of COMPOSE on the CMU Panoptic dataset [JLT<sup>+</sup>15] against the number of people in the scene. With the standard 5-camera setup, the solver converges with an average runtime of 22.7ms per frame on a workstation with an Intel Xeon Gold 6336Y CPU @ 2.40 GHz.

## 5 Concluding Remarks

We presented COMPOSE, a novel graph formulation for modeling 3D multi-view multi-human pose estimation. By representing 2D body joint detections as local graphs within a global multi-view hypergraph, our approach effectively captures the complex relationships inherent in multi-view settings. With this formulation, we cast the correspondence search as a weighted exact cover problem, which we show is efficiently solvable using Integer Linear Programming (ILP). Extensive experiments on multiple benchmarks demonstrate that our method achieves state-of-the-art performance in 3D multi-view multi-human pose estimation, even surpassing learning-based approaches in several settings. Our method’s performance highlights the potential of higher-order graph formulations for multi-view 3D pose estimation tasks, with potential to improve downstream applications such as tracking [WHNB25] and workflow recognition [BCH<sup>+</sup>23].

**Acknowledgements.** The authors gratefully acknowledge the computational resources of the LRZ AI service infrastructure provided by the Leibniz Supercomputing Center (LRZ) and Munich Center for Machine Learning (MCML), funded by the German Federal Ministry of Education and Research (BMBF) and Bavarian State Ministry of Science and the Arts (StMWK). T. W., L.B., and N.N. acknowledge support from the TUM Global Incentive Fund project TOPO-OR. T. B. was supported by a UKRI Future Leaders Fellowship (MR/Y018818/1).

## References

- [BAA<sup>+</sup>14] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3D Pictorial Structures for Multiple Human Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1669–1676, 2014.
- [BCH<sup>+</sup>23] Lennart Bastian, Tobias Czempiel, Christian Heiliger, Konrad Karcz, Ulrich Eck, Benjamin Busam, and Nassir Navab. Know your sensors—a modality study for surgical action classification. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 11(4):1113–1121, 2023.
- [BS19] Tolga Birdal and Umut Simsekli. Probabilistic permutation synchronization using the riemannian structure of the birkhoff polytope. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11105–11116, 2019.
- [CGD25] Aviral Chharia, Wenbo Gou, and Haoye Dong. Mv-ssm: Multi-view state space modeling for 3d human pose estimation. pages 11590–11599, 2025.
- [CKJ23] Rohan Choudhury, Kris M. Kitani, and László A. Jeni. TEMPO: Efficient Multi-View Pose Estimation, Tracking, and Forecasting. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14704–14714. IEEE, 2023.
- [CSMP25] Keqi Chen, Vinkle Srivastav, Didier Mutter, and Nicolas Padoy. Learning from synchronization: Self-supervised uncalibrated multi-view person association in challenging scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24419–24428, 2025.
- [CT25] Bo-Han Chen and Chia-chi Tsai. 3DSA: Multi-view 3D Human Pose Estimation With 3D Space Attention Mechanisms. In *Computer Vision – ECCV 2024*, volume 15085, pages 323–339. Springer Nature Switzerland, 2025.
- [DFJ<sup>+</sup>22] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and Robust Multi-Person 3D Pose Estimation and Tracking From Multiple Views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6981–6992, 2022.
- [ENKS18] Sara Ershadi-Nasab, Erfan Noury, Shohreh Kasaei, and Esmaeil Sanaei. Multiple human 3D pose estimation from multiview images. *Multimedia Tools and Applications*, 77(12):15573–15601, 2018.
- [GS08] Michael A Goodrich and Alan C Schultz. Human–robot interaction: a survey. *Foundations and trends® in human–computer interaction*, 1(3):203–275, 2008.
- [HSZW21] Buzhen Huang, Yuan Shu, Tianshu Zhang, and Yangang Wang. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *2021 International Conference on 3D Vision (3DV)*, pages 710–720. IEEE, 2021.
- [HWB<sup>+</sup>21] Jiahui Huang, He Wang, Tolga Birdal, Minhyuk Sung, Federica Arrigoni, Shi-Min Hu, and Leonidas J Guibas. Multibodysync: Multi-body segmentation and motion estimation via 3d scan synchronization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7108–7118, 2021.
- [HWY<sup>+</sup>22] Ruize Han, Yun Wang, Haomin Yan, Wei Feng, and Song Wang. Multi-view multi-human association with deep assignment network. *IEEE Transactions on Image Processing*, 31:1830–1840, 2022.
- [HZ03] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.
- [IBLM19] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable Triangulation of Human Pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2019.

- [JLT<sup>+</sup>15] 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 *The IEEE International Conference on Computer Vision (ICCV)*, 2015.
- [Kar09] Richard M Karp. Reducibility among combinatorial problems. In *50 Years of Integer Programming 1958-2008: from the Early Years to the State-of-the-Art*, pages 219–241. Springer, 2009.
- [LD60] Ailsa H. Land and Alison G. Doig. An automatic method of solving discrete programming problems. *Econometrica*, 28(3):497–520, 1960.
- [LL21] Jiahao Lin and Gim Hee Lee. Multi-View Multi-Person 3D Pose Estimation with Plane Sweep Stereo. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11881–11890. IEEE, 2021.
- [LXK<sup>+</sup>24] Yu-Jhe Li, Yan Xu, Rawal Khirodkar, Jinyung Park, and Kris Kitani. Multi-person 3d pose estimation from multi-view uncalibrated depth cameras. *arXiv preprint arXiv:2401.15616*, 2024.
- [LZ25] Yang Liu and Zhiyong Zhang. DSP: Dense-Sparse Parallel Networks for Self-supervised 3D Multi-person Pose Estimation from Multiple Views. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4629–4638. ACM, 2025.
- [LZW<sup>+</sup>24] Ziwei Liao, Jialiang Zhu, Chunyu Wang, Han Hu, and Steven L. Waslander. Multiple View Geometry Transformers for 3D Human Pose Estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 708–717. IEEE, 2024.
- [MOD11] Stuart Mitchell, Michael OSullivan, and Iain Dunning. Pulp: a linear programming toolkit for python. *The University of Auckland, Auckland, New Zealand*, 65:25, 2011.
- [NOT25] Ana Filipa Rodrigues Nogueira, Hélder P. Oliveira, and Luís F. Teixeira. Markerless multi-view 3D human pose estimation: A survey. *Image and Vision Computing*, 155:105437, 2025.
- [PGS19] Aleksis Pirinen, Erik Gärtner, and Cristian Sminchisescu. Domes to Drones: Self-Supervised Active Triangulation for 3D Human Pose Reconstruction. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [SCP24] Vinkle Srivastav, Keqi Chen, and Nicolas Padoy. SelfPose3d: Self-Supervised Multi-Person Multi-View 3d Pose Estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2502–2512. IEEE, 2024.
- [TWZ20] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. VoxelPose: Towards Multi-camera 3D Human Pose Estimation in Wild Environment. In *Computer Vision – ECCV 2020*, pages 197–212. Springer International Publishing, 2020.
- [WBC<sup>+</sup>25] Tony Danjun Wang, Lennart Bastian, Tobias Czempiel, Christian Heiliger, and Nassir Navab. Beyond role-based surgical domain modeling: Generalizable re-identification in the operating room. *Medical Image Analysis*, page 103687, 2025.
- [WHNB25] Tony Danjun Wang, Christian Heiliger, Nassir Navab, and Lennart Bastian. Trackor: Towards personalized intelligent operating rooms through robust tracking. In *International Workshop on Collaborative Intelligence and Autonomy in Image-Guided Surgery*, pages 53–63. Springer, 2025.
- [WJL<sup>+</sup>21] Size Wu, Sheng Jin, Wentao Liu, Lei Bai, Chen Qian, Dong Liu, and Wanli Ouyang. Graph-Based 3D Multi-Person Pose Estimation Using Multi-View Images. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11128–11137. IEEE, 2021.

- [WZC<sup>+</sup>21] Tao Wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. Direct Multi-view Multi-person 3D Pose Estimation. In *Advances in Neural Information Processing Systems*, volume 34, pages 13153–13164. Curran Associates, Inc., 2021.
- [XZZT22] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.
- [YZW<sup>+</sup>22] Hang Ye, Wentao Zhu, Chunyu Wang, Ruijie Wu, and Yizhou Wang. Faster VoxelPose: Real-time 3D Human Pose Estimation by Orthographic Projection. In *Computer Vision – ECCV 2022*, pages 142–159. Springer Nature Switzerland, 2022.
- [ZAY<sup>+</sup>20] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4D Association Graph for Realtime Multi-Person Motion Capture Using Multiple Video Cameras. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1321–1330. IEEE, 2020.
- [ZSW<sup>+</sup>22] Zhize Zhou, Qing Shuai, Yize Wang, Qi Fang, Xiaopeng Ji, Fashuai Li, Hujun Bao, and Xiaowei Zhou. QuickPose: Real-time Multi-view Multi-person Pose Estimation in Crowded Scenes. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH ’22, pages 1–9. Association for Computing Machinery, 2022.
- [ZWC<sup>+</sup>23] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM computing surveys*, 56(1):1–37, 2023.