3D Pose Analysis: Predict, Match & Search

Megan Dass
Stanford University
mdass9@stanford.edu

Aman Kansal
Stanford University
amkansal@stanford.edu

Abstract

The evolving landscape of technology increasingly emphasizes dynamic data modalities, such as imagery and speech, necessitating innovative advancements in search infrastructure to effectively manage these formats. Our project focuses on a specialized area within image-based search: pose matching. Traditional image search methods predominantly utilize 2D image data, leading to inaccuracies when representing inherently three-dimensional poses. Addressing this limitation, our project explores advanced 3D pose estimation techniques using state-of-the-art models such as DiffPose for more accurate and reliable pose matching. DiffPose excels in transforming 2D heatmaps into plausible 3D pose predictions. Additionally, traditional pose comparison methods often involve computationally expensive 3D point cloud warping. To overcome this, we experiment with and show improvements in the POEM model that embeds 2D images directly into 3D pose representations and also develop a simple yet effective method to derive 3D pose embeddings from 3D point clouds, achieving efficiency without sacrificing accuracy. Our project also curates an open-source dataset featuring humans in identical poses across varied environments to aid in testing and refining pose matching algorithms. This holistic approach not hopes to enhance user experience in applications ranging from fitness to professional modeling.

1. Introduction

Pose matching is an increasingly crucial domain in today's tech landscape as we transition from text-based modalities to more dynamic forms such as speech and imagery. This shift necessitates a parallel evolution in search infrastructure to accommodate and fully leverage these new forms of data. Within the vast realm of image-based search, our project addresses a particularly specialized yet significant challenge: pose matching.

Currently, the digital marketplace is replete with applications designed to assist users in a variety of ways, from choosing the right hairstyle to mastering gym exercises, and even emulating celebrity photo poses. These applications rely on traditional image search techniques, which primarily utilize 2D image data. However, from our studies in CS 231A, it has become evident that such problems are inherently three-dimensional. Relying solely on 2D data can introduce substantial inaccuracies, as the true essence of the pose can significantly deviate in the translation from 3D reality to 2D representation.

Our course project aims to refine the approach to modeling poses, focusing particularly on helping amateurs and actors achieve more accurate poses by comparing their efforts with those of established figures known for their proficiency in specific poses. This comparison utilizes sophisticated image search algorithms where images are processed, converted into embeddings, and matched using distance metrics such as cosine similarity or L2 norm. Recognizing the limitations inherent in 2D images and videos for capturing true 3D forms, our project explores advanced 3D pose estimation techniques. By tapping into strategic points within existing state-of-the-art 3D pose prediction models, we aim to derive accurate and directly comparable 3D pose representations.

Traditional methods of pose comparison use 3D point clouds with algorithms where the source point cloud undergoes warping or non-rigid transformation to match the target point cloud, known as Non-rigid Registration. While accurate, these methods are inefficient. We explore and improve POEM (Google's 2D image to 3D pose embedding model) using transfer learning and propose a simple yet efficient way to derive 3D pose embeddings directly from the 3D point cloud, performing comparably to deep learning models.

A significant gap in this field is the lack of accessible open-source datasets featuring humans in identical poses across varied environments. Such datasets are essential for rigorously testing and validating pose matching algorithms. To address this, our project includes an initiative to curate a comprehensive dataset that groups together images of humans in similar poses but in different settings, thereby providing a robust resource for assessing the effectiveness of pose matching techniques.

The code for our experiments and to recreate our results can be accessed here: https://github.com/kansalaman/3d-pose-estimation-and-matching.

2. Background & Related Work

The field of computer vision is increasingly tasked with handling dynamic data, including images representing 3D objects and actions. Traditional image search and pose matching methods often struggle with accurately representing the inherent 3D nature of poses within these images. This limitation motivates the need for more advanced techniques in pose matching for various applications.

Non-rigid registration techniques offer a solution for aligning non-deformable shapes in 3D space. However, these methods can be computationally expensive, in having to duplicate computation for every pair of data leading to a quadratic complexity [2]. The Iterative Closest Point (ICP) algorithm provides a versatile and efficient means of aligning 3D shapes by iteratively minimizing the mean-square distance between corresponding points. While effective for rigid objects and initial estimations for non-rigid registration tasks, it is not scale invariant, which is not ideal for the pose-matching use case [1].

Among the challenges in 3D pose estimation is the transformation from 2D keypoints to a plausible 3D representation. Sun et al. introduce Pr-VIPE, also known as POEM, a method for learning probabilistic view-invariant embeddings solely from 2D human pose keypoints. By leveraging probabilistic embeddings, POEM effectively captures the inherent ambiguity in 2D poses, enhancing tasks such as cross-view pose retrieval and action recognition. This approach showcases the importance of robust representation learning in addressing the challenges of 3D pose estimation from 2D data [4].

Recent advancements in pose estimation address this challenge. DiffPose stands out as a state-of-the-art conditional diffusion model method for predicting multiple 3D pose hypotheses from a single 2D image. This approach tackles the inherent ambiguity in pose estimation and avoids overconfident single-pose predictions [3].

Our project proposes a novel approach to 3D pose embedding generation by leveraging the strengths of existing methods. Specifically, we introduce a method where we utilize the 3D keypoints generated by DiffPose as input to the POEM model, thereby enabling the transformation from 3D pose to embedding. By combining elements from the surveyed techniques and recent advancements, our method aims to provide a simple, yet accurate solution for enhancing pose matching accuracy.

3. Approach

In our approach, we use the DiffPose model for 3D pose prediction and the POEM model for 3D pose matching. The combination of these models leverages state-of-the-art techniques to produce accurate and robust 3D pose estimations.

3.1. DiffPose Model

The DiffPose model comprises two main components: a 2D pose prediction model and a diffusion model. The 2D pose prediction model is pretrained and fine-tuned jointly with the diffusion model. Specifically, the HRNet model is employed for 2D pose predictions.

3.1.1 The Forward Process

The forward process in the DiffPose model involves transforming an initial data point x_0 by adding Gaussian noise over several steps, modeled by a Markov chain. The objective is to convert the data point to a Gaussian distribution $\mathcal{N}(0, I)$ through the following process:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

Here:

- x_t is the data point at step t
- β_t is the noise variance schedule parameter at step t
- I is the identity matrix

3.1.2 The Reverse Process

The reverse process aims to iteratively restore the original data by modeling each step as a Gaussian distribution. This helps in effectively undoing the noise added during the forward process. For each step t, the transition from x_t to x_{t-1} is modeled as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, c), \Sigma_\theta(x_t, c))$$

where:

- $\mu_\theta(x_t, t)$ is the mean of the Gaussian distribution
- $\Sigma_\theta(x_t, t)$ is the variance of the Gaussian distribution
- c is the 2D joint representation

3.1.3 2D Joint Representations

The conditioning input c is derived from the 2D pose prediction model, which outputs heatmaps for each joint. These heatmaps are transformed using a transformer network and subsequently fed into a feedforward neural network to produce a mixed inter-joint representation:

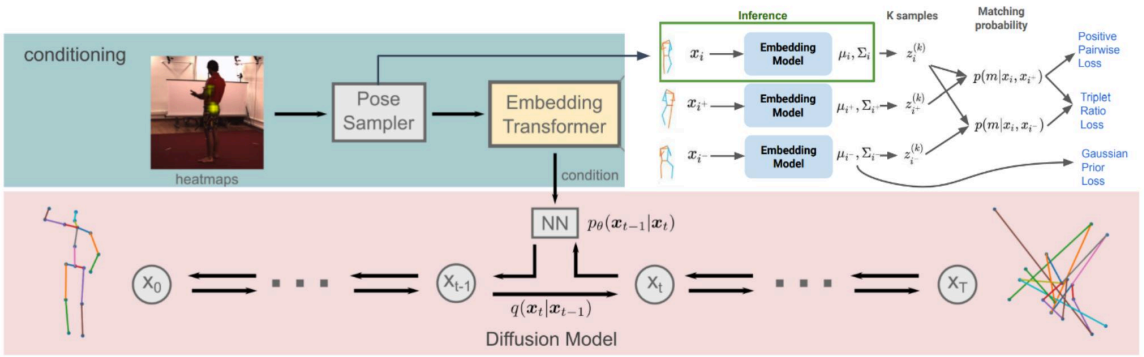

Figure 1. Model architecture: Diffpose + POEM

$$c = \text{Feedforward}(\text{Transformer}(\text{Heatmaps}))$$

This conditioning input c is crucial for the diffusion model, which is trained jointly with the 2D pose prediction model. The joint training allows the models to enhance each other, where the diffusion model refines the 2D pose predictions and the improved 2D predictions, in turn, enhance the 3D pose estimation through a shared backpropagation signal.

3.2. POEM Model

The POEM (Pose Embedding) model is used for 3D pose matching. It defines an embedding space where similar 3D poses are close together, and dissimilar ones are far apart. This approach ensures robust pose matching and helps in improving the accuracy of 3D pose predictions.

3.2.1 Matching Definition

Two 3D poses are defined as matching if they are visually similar, regardless of the viewpoint. To formalize this,

$$m_{ij} = \begin{cases} 1, & \text{if similar}(y_i, y_j) \\ 0, & \text{otherwise} \end{cases}$$

3.2.2 Triplet Ratio Loss

The core of the POEM model is the triplet ratio loss, which ensures that similar 3D poses (positive pairs) have higher matching probabilities than dissimilar ones (negative pairs). For a given triplet (x_i, x_i^+, x_i^-) with $m_{i,i^+} > m_{i,i^-}$, the objective is:

$$\frac{p(m|z_i, z_i^+)}{p(m|z_i, z_i^-)} \geq \beta$$

where:

- $z = f(x)$ is the embedding of the input 2D pose x.
- $\beta > 1$ is the ratio representing the matching probability of a similar 3D pose pair compared to a dissimilar pair.

Taking the negative logarithm of both sides, we get:

$$-\log p(m|z_i, z_i^+) + \log p(m|z_i, z_i^-) \leq -\log \beta$$

Given a batch size N, the triplet ratio loss L_{ratio} is defined as:

$$L_{\text{ratio}} = \sum_{i=1}^{N} \max(0, D_m(z_i, z_i^+) - D_m(z_i, z_i^-) + \alpha)$$

where:

- D_m is the distance measure: $-\log p(m|...)$
- α is a margin parameter: $-\log \beta$

Note that DiffPose fine-tunes the 2D pose detector in end-to-end training. We use this supervision to benefit POEM by inputting fine-tuned 2D pose representations for 3D tasks instead of vanilla 2D pose representations.

3.3. Similarity Functions

To compare the embeddings generated by POEM, we use both L2 and cosine similarity metrics, with cosine similarity performing better in our experiments.

3.3.1 Iterative Closest Point (ICP) Algorithm

We explore the ICP algorithm, which iteratively computes the translation and rotation matrices between a source point cloud and a set of target point clouds to best fit the two

before calculating the L2 distance. The ICP algorithm involves the following steps:

$$\mathbf{H}^{(k)} = \sum_{i=1}^{N} (\mathbf{p}_i - \bar{\mathbf{p}}^{(k)})(\mathbf{q}_i^{(k)} - \bar{\mathbf{q}}^{(k)})^T \quad (1)$$

where:

- $\mathbf{H}^{(k)}$ is the cross-covariance matrix at iteration k.
- $\mathbf{p}_i$ and $\mathbf{q}_i^{(k)}$ are the 3D keypoints from the source and target point clouds, respectively.
- $\bar{\mathbf{p}}^{(k)}$ and $\bar{\mathbf{q}}^{(k)}$ are the centroids of the source and target point clouds.

Using Singular Value Decomposition (SVD), we decompose $\mathbf{H}^{(k)}$:

$$\mathbf{H}^{(k)} = \mathbf{U}^{(k)} \mathbf{\Sigma}^{(k)} (\mathbf{V}^{(k)})^T \quad (2)$$

The translation and rotation matrices are updated as follows:

$$\mathbf{t}^{(k+1)} = \bar{\mathbf{q}}^{(k)} - \mathbf{R}^{(k+1)} \bar{\mathbf{p}}^{(k)} \quad (3)$$

$$\mathbf{R}^{(k+1)} = \mathbf{V}^{(k)} (\mathbf{U}^{(k)})^T \quad (4)$$

3.3.2 Interjoint Distance Embedding

We also propose a simple yet effective embedding method: interjoint distances. This method represents the normalized pairwise distances between the 3D coordinates of all 16 keypoints, resulting in a 120-dimensional embedding. The interjoint distances are computed as follows:

$$d_{ij} = ||\mathbf{p}_i - \mathbf{p}_j||_2 = \sqrt{(p_i^x - p_j^x)^2 + (p_i^y - p_j^y)^2 + (p_i^z - p_j^z)^2} \quad (5)$$

$$\mathbf{d} = [d_{12}, d_{13}, \ldots, d_{1N}, d_{23}, \ldots, d_{2N}, \ldots, d_{(N-1)N}] \quad (6)$$

$$\mathbf{d}_{\text{normalized}} = \frac{\mathbf{d}}{||\mathbf{d}||_2} \quad (7)$$

where:

- d_{ij} is the L2 distance between keypoints i and j.
- $\mathbf{p}_i$ and $\mathbf{p}_j$ are the 3D coordinates of the keypoints.
- $\mathbf{d}$ is the vector of all pairwise distances.
- $\mathbf{d}_{\text{normalized}}$ is the normalized vector of interjoint distances.

This embedding provides a normalized representation of the spatial relationships between the keypoints, enhancing the performance of the POEM model.

Pose Label	Number of Images
Arms Raised	52
Crossed Arms	70
Hands Clasped	29
Hands in Pocket	52
Jumping in Air	41
Looking Over Shoulder	59
One Hand on Hip	35
Sitting on Ground	70
Total	**408**

Table 1. Dataset Breakdown per Pose Label.

4. Experiments & Analysis

4.1. Dataset

We curated a custom dataset comprising eight distinct classes of human poses sourced from the Google Image API [1]. These classes were selected to encompass a broad spectrum of poses, ranging from common everyday gestures to more specialized actions. Moreover, we intentionally included poses that exhibited similarities, such as "hands in pocket" versus "one hand on hip," to assess the robustness of our system in handling occlusions and variations in pose presentation.

Prior to inclusion in the dataset, we filtered the API results to ensure uniformity in pose presentation and environmental context for the reliability and consistency of the dataset.

Each pose class within the dataset comprises of a varying number of images, with the range spanning from 29 to 70 images per class. In total, our curated dataset consists of 408 images distributed across the eight pose classes.

4.2. Results

We compare the following methods for 3D pose prediction and matching:

1. **POEM with Cosine Similarity:** Using POEM to get 3D pose embeddings and comparing them using cosine similarity.
2. **POEM with DiffPose's 2D Pose Generator:** Using POEM with DiffPose's finetuned 2D pose generator for getting 3D pose embeddings and comparing using cosine similarity.
3. **ICP on DiffPose's 3D Pose Predictions:** Applying the Iterative Closest Point (ICP) algorithm on DiffPose's 3D pose predictions.

[1] https://developers.google.com/custom-search/v1/overview

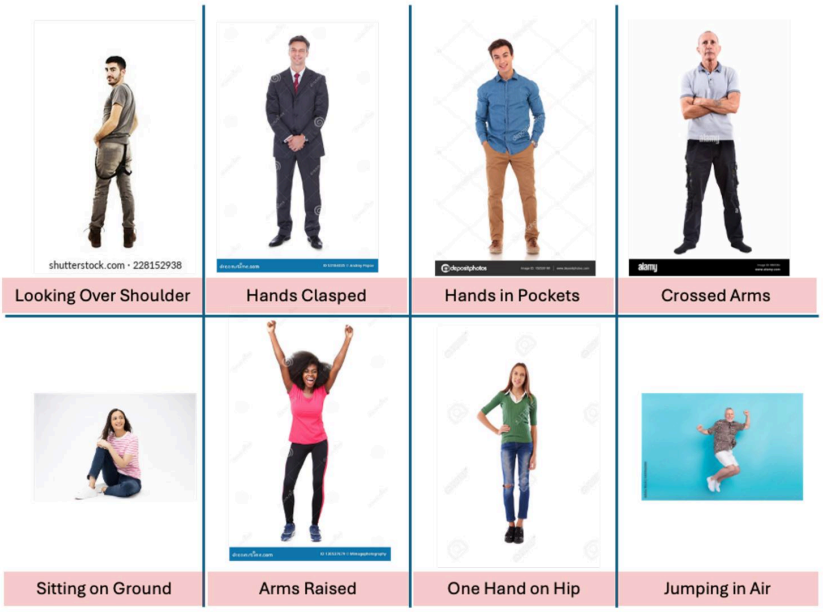

Figure 2. Sample image for each pose label in curated dataset.

4. **Cosine Distance on Interjoint Embeddings:** Calculating cosine distance between interjoint embeddings constructed from DiffPose's 3D pose predictions.

Evaluation Metrics: We evaluate the performance of these methods using accuracy, precision, and recall. The definitions are as follows:

- **Accuracy:** The proportion of correctly predicted poses out of all predictions.
- **Precision:** The proportion of true positive predictions out of all positive predictions.
- **Recall:** The proportion of true positive predictions out of all actual positives.

The metrics for each model are calculated as a weighted average of the pose-level metrics. The results are presented in Figure 5.

Pose Recommendations: To illustrate the effectiveness of the different models, we input two real-life images and show the corresponding pose recommendations by different models in Figure 3.

Confusion Matrix Comparison: Finally, we compare the confusion matrices of our best-performing model and the interjoint distances method in Figure 4.

Analysis of Results:

- **Performance of ICP Method:** The ICP method was unique among the compared approaches in that it utilized multiple prediction hypotheses produced by DiffPose. This approach aimed to iteratively refine the alignment between the source and target 3D pose predictions. Despite the theoretical advantage of using multiple hypotheses, the ICP method performed the worst among all methods. This suggests that the additional complexity introduced by considering multiple hypotheses did not translate into improved performance, possibly due to the challenge of correctly aligning noisy or imperfect pose predictions.
- **Invariance Properties of POEM and Interjoint Distances:** The POEM model (both the standard version and the variant using DiffPose's finetuned 2D pose generator) and the Interjoint Distances method are designed to be invariant to scale, translation, and rotation. These properties are critical for accurately comparing poses regardless of variations in size, position, or orientation. The substantial performance gap observed between these methods and the ICP method underscores the importance of scale invariance in achieving robust pose matching. Specifically, the inability of ICP to handle scale variations effectively likely contributed to its lower performance.
- **Dimensionality and Performance of Embeddings:** POEM embeddings are compact, 16-dimensional vec-

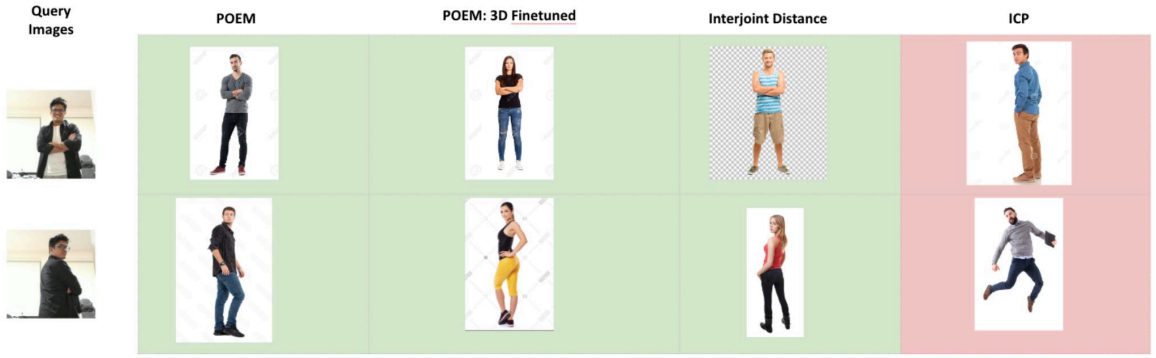

Figure 3. Pose recommendations by different models for two real-life images.

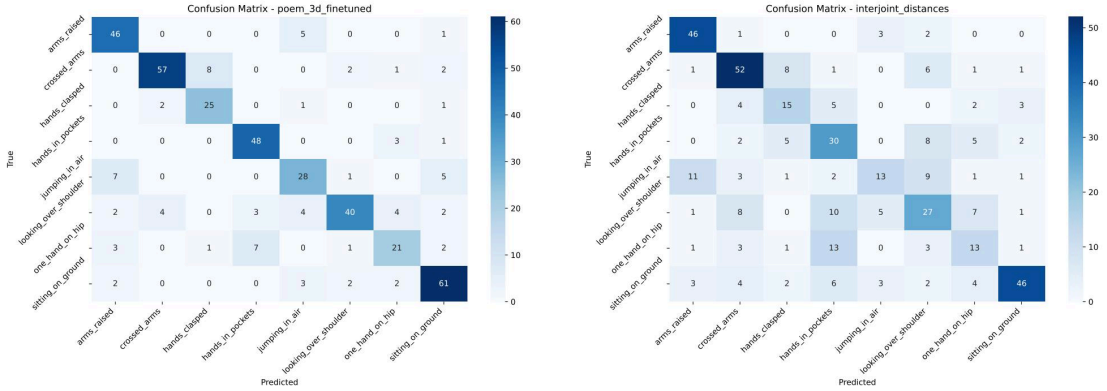

Figure 4. Comparison of confusion matrices for the best-performing model and interjoint distances method. (This image is vectorized, zoom in for better visibility)

tors, whereas Interjoint Distance embeddings are more extensive, 120-dimensional vectors. Despite the higher dimensionality of Interjoint Distance embeddings, which theoretically could capture more detailed pose information, POEM embeddings consistently performed better across all metrics. This indicates that POEM's embedding space is more effective at representing the essential characteristics of 3D poses, leading to superior performance even with a lower-dimensional representation. Also quick SVD analysis of Interjoint Distance Embeddings showed the matrix to be comparatively low rank but with all singular values non zero pointing towards a comparatively high amount of noise in the embeddings.

- **Handling of Occlusions by POEM:** One of the significant advantages of POEM is its ability to handle occlusions effectively. Occlusions, where parts of the body are hidden from view, pose a significant challenge for pose estimation models. POEM demonstrates robustness in such scenarios, leading to better performance compared to Interjoint Distances. Common confusion cases for Interjoint Distances, which POEM handles better, include:

  – **"Hands in Pocket" vs. "Hands on Hips":** These poses can appear similar when parts of the body are occluded, but POEM can distinguish them more accurately.

  – **"Crossed Arms" vs. "Looking Over Shoulder":** A large number of "Looking Over Shoulder" images also have subjects with crossed arms but which are occluded.

5. Conclusion

Traditional pose matching methods often relied on ad-hoc comparisons of two images, which, while accurate were

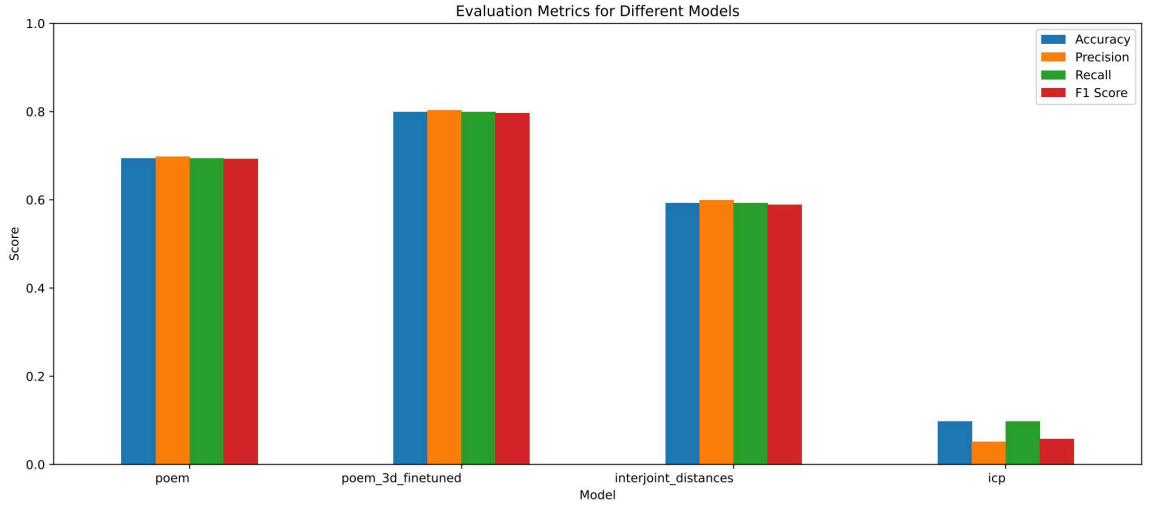

Figure 5. Comparison of accuracy, precision, and recall for different approaches.

fairly inefficient. However still, to experiment, we implemented the Iterative Closest Point (ICP) algorithm to leverage multiple prediction hypotheses produced by DiffPose. However, ICP struggled significantly with scale invariance, leading to poor performance compared to other methods.

We validated the effectiveness of the POEM model for pose matching, demonstrating that it produces improved embeddings through the use of transfer learning. The incorporation of DiffPose's finetuned 2D pose generator further enhanced the quality of the embeddings, resulting in superior performance.

Additionally, we proposed an efficient method for direct 3D pose embedding using interjoint distances. Although this method generated higher-dimensional embeddings, POEM's lower-dimensional embeddings consistently outperformed it, highlighting the efficacy of POEM's embedding space.

For future work, we aim to improve the robustness of ICP to scale variations by incorporating concepts from C231A. Enhancing ICP in this manner could potentially bridge the performance gap observed in our study and make it a more viable option for 3D pose matching.

References

[1] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992. 2

[2] Bailin Deng, Yuxin Yao, Roberto M Dyke, and Juyong Zhang. A survey of non-rigid 3d registration. In *Computer Graphics Forum*, volume 41, pages 559–589. Wiley Online Library, 2022. 2

[3] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15977–15987, 2023. 2

[4] Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 53–70. Springer, 2020. 2