Polar Coordinate-Based 2D Pose Prior with Neural Distance Field

Qi Gan[1]   Sao Mai Nguyen[2]   Eric Fenaux[3]   Stephan Clémençon[1]

Mounim El Yacoubi[4]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2]U2IS, ENSTA Paris, Institut Polytechnique de Paris, France
[3]Ef-e-science, France
[4]SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, France

{qi.gan, stephan.clemencon}@telecom-paris.fr,
{nguyensmai, eric.fenaux}@gmail.com, mounim.el_yacoubi@telecom-sudparis.eu

Abstract

Human pose capture is essential for sports analysis, enabling precise evaluation of athletes' movements. While deep learning-based human pose estimation (HPE) models from RGB videos have achieved impressive performance on public datasets, their effectiveness in real-world sports scenarios is often hindered by motion blur, occlusions, and domain shifts across different pose representations. Fine-tuning these models can partially alleviate such challenges but typically requires large-scale annotated data and still struggles to generalize across diverse sports environments. To address these limitations, we propose a 2D pose prior-guided refinement approach based on Neural Distance Fields (NDF). Unlike existing approaches that rely solely on angular representations of human poses, we introduce a polar coordinate-based representation that explicitly incorporates joint connection lengths, enabling a more accurate correction of erroneous pose estimations. Additionally, we define a novel non-geodesic distance metric that separates angular and radial discrepancies, which we demonstrate is better suited for polar representations than traditional geodesic distances. To mitigate data scarcity, we develop a gradient-based batch-projection augmentation strategy, which synthesizes realistic pose samples through iterative refinement. Our method is evaluated on a long jump dataset, demonstrating its ability to improve 2D pose estimation across multiple pose representations, making it robust across different domains. Experimental results show that our approach enhances pose plausibility while requiring only limited training data. Code is available at: https://github.com/QGAN2019/polar-NDF.

1. Introduction

Human pose capture plays a vital role in sports analysis, particularly for the quantitative evaluation of athletic performance [2]. By analyzing pose sequences, the technical intricacies of an athlete's movements can be measured and assessed with high precision [9, 19, 28]. Despite its importance, accurately capturing human pose in sports scenarios remains a significant challenge.

Traditionally, capturing human movement requires annotating human poses from videos frame-by-frame, which is time-consuming and costly, or dedicated motion capture systems requiring athletes to wear sensors, suits, or physical markers. However, many sports applications demand fast, real-time feedback and non-intrusive pose acquisition. Recent advances in computer vision offer a promising alternative through deep neural network (DNN)-based human pose estimation [3, 7, 23, 27]. Despite achieving human-level accuracy on public datasets [16], these models often produce subpar results when applied to real-world sports videos. This performance gap stems from domain differences between the curated training datasets and the complex nature of sports footage. Sports videos commonly exhibit low frame rates, significant motion blur, dynamic lighting conditions, and frequent multi-person occlusions. One strategy to address these issues is fine-tuning 2D Human Pose Estimation (HPE) models on domain-specific data. For instance, Ludwig *et al.* [17] fine-tuned HRNet [23] using a self-supervised approach. However, this method is limited by the diversity of visual contexts in sports videos, including variations in background, lighting, uniforms, and camera angles—even within the same sport. Moreover, inconsistencies in pose representations across models further complicate result transferability [3, 7, 21]. A more robust
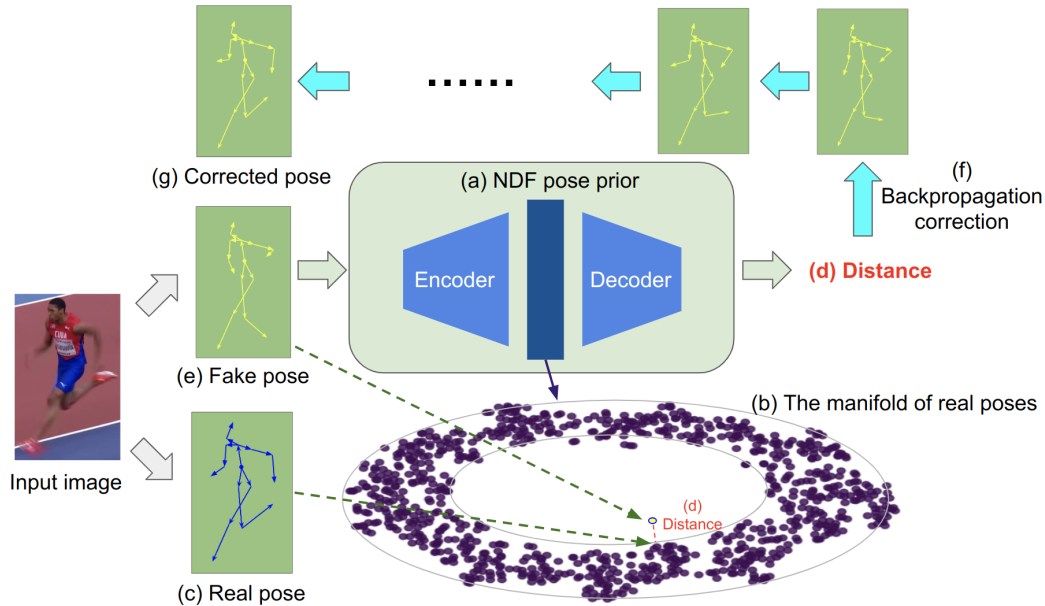
Figure 1. Overview of the method. We leverage a Neural Distance Field (NDF) network to learn the underlying pose prior manifold from real human poses. The (a) NDF network learns (b) the zero-level-set manifold (represented as deep blue points and visualized using t-SNE [25]) of (c) real poses. The system estimates the (d) distance of (e) a fake pose in the embedded space (the yellow point) to the manifold. By performing several iterations of (f) backpropagation with the parameters of NDF fixed, the input fake pose is corrected, outputing (g) a corrected pose.

alternative involves post-correcting estimated poses using motion and structural constraints, such as temporal continuity [29, 30], which has proven effective for correcting limb joint swaps. However, while temporal smoothing can fix local inconsistencies, it is insufficient for correcting broader detection errors. A more effective approach to improving HPE results is to integrate prior knowledge of human pose into a dedicated pose prior model. While recent work on pose priors has predominantly focused on 3D human poses [1, 24, 26], modeling 2D pose priors presents unique challenges. This is primarily because a single 3D pose can be projected into multiple valid 2D configurations, depending on the viewing angle or camera perspective. POST [21] successfully applied a 2D pose prior with Neural Distance Fields (NDF) [5] for domain adaptation in 2D pose estimation using angular representations, where joint connection directions are encoded via their cosine and sine values. However, this angular representation ignores the lengths of joint connections (i.e., limb lengths), which are critical for evaluating the plausibility of a pose. A pose that is valid in terms of joint angles may become physically implausible if the limb lengths are altered. Therefore, angular 2D pose priors may face limitations when applied directly to certain HPE scenarios. Moreover, training an NDF-based pose prior typically requires large-scale datasets [21, 24].

To enable accurate and robust estimation of sports poses with limited training data, we propose a 2D pose prior-guided refinement method. Our approach extends prior work on neural distance field (NDF)-based pose priors [21, 24], which model the assumption that plausible human poses reside on a zero-level manifold, meaning all valid poses have zero distance to the manifold in the embedded space. In contrast, implausible poses lie outside this manifold, resulting in positive distances that reflect their deviation from physically realistic configurations. To enable efficient correction of erroneous estimations, we introduce a novel pose representation and an accompanying distance metric. Unlike POST [21], our method employs a polar coordinate-based pose representation that explicitly incorporates joint connection lengths, enabling the pose prior to more effectively refine 2D estimations. In conjunction with this representation, we propose a non-geodesic distance metric that separately captures angular and radial discrepancies. Leveraging this metric and representation, we further propose a gradient-based pose augmentation strategy that allows the NDF model to be trained with limited data. By backpropagating through arbitrary batches of poses, our method generates realistic synthetic samples that improve training efficiency. We validate our approach on a long jump dataset introduced by Gan *et al.* [9]. The contributions of this work include:

- We propose a polar coordinate-based pose representation and a corresponding distance metric for NDF-based 2D human pose priors, along with a novel training scheme.

This formulation enables the pose prior to refine both angular and radial errors in 2D pose estimations.

- We introduce a gradient-based data augmentation method that synthesizes new pose samples during training. This strategy allows the 2D NDF prior to be effectively learned from only a few video clips.
- We experimentally demonstrate that the learned 2D pose prior improves erroneous pose estimations across multiple 2D pose representations, enhancing robustness across different pose domains.

2. Related works

2.1. Human pose priors

Recent studies on human pose priors generally fall into two main categories. The first category involves Variational Autoencoder (VAE)-based approaches. For example, VPose [20] employs a variational prior to model human body poses, while HuMoR [22] utilizes a conditional VAE to capture both temporal dynamics and body shape. The second category leverages neural fields to represent human pose priors. Pose-NDF [24] models 3D human poses using a Neural Distance Field (NDF) [5], while H-NeRF [26] learns 3D pose priors using Neural Radiance Fields (NeRF) [18]. However, most existing methods focus on modeling 3D pose priors. In contrast, learning priors for 2D poses presents significantly greater challenges, as a single 3D human pose can project to many different 2D configurations depending on the camera viewpoint. Moreover, unlike 3D poses where joint connection lengths are fixed, 2D joint lengths vary continuously from zero to their full projected length, introducing additional variability. To the best of our knowledge, the only existing work that models 2D human pose priors is POST [21], which employs a NDF. Their method represents 2D human poses solely through the orientations of joint connections, neglecting connection lengths. This omission is potentially problematic, as joint length is a critical factor in biomechanical plausibility, modifying connection lengths while preserving orientations can result in implausible poses. Furthermore, NDF-based priors typically require large-scale datasets for training [21, 24], limiting their applicability in data-scarce scenarios. Our approach adopts a polar coordinate-based representation to learn 2D pose priors using NDF, enabling the model to correct errors in both joint connection orientations and lengths. Additionally, we introduce a gradient projection-based data augmentation strategy that facilitates effective NDF learning even with limited training data.

2.2. Sports pose estimation

Applying HPE models directly to sports videos often yields suboptimal results due to the domain gap between everyday human poses and those specific to sports, as well as the
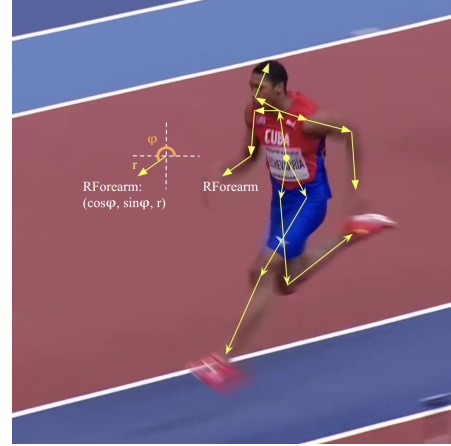

Figure 2. The 2D human pose representation used in this work is based on polar coordinates and consists of 15 joint connection vectors. Each vector is encoded by the cosine and sine of its orientation φ, along with its length (see the example of the 'RForearm' in the figure).

degraded video quality caused by motion blur and occlusions. Two main approaches have been proposed to mitigate these issues. The first is model fine-tuning. To address the challenge of limited annotated data, Ludwig *et al*. [17] introduced two self-supervised learning schemes to enhance pose estimation performance, which were evaluated on long jump sequences. The second approach involves correcting estimated poses. Fastovets *et al*. [8] proposed an interactive, model-based generative framework for pose estimation in broadcast sports footage. Zecha *et al*. [29, 30] refined swimming pose estimates using temporal consistency. Dittakavi *et al*. [6] introduced an explainable pose correction system based on an angle-likelihood mechanism, applied to yoga pose analysis. In our work, we learn a 2D pose prior and address the data scarcity problem through a specialized gradient-based data augmentation strategy.

3. Methodology

Our objective is to learn a 2D pose prior to correct erroneous pose estimations from sports videos (mostly broadcast competition videos) with a small amount of data. To achieve this goal, we propose a 2D polar coordinate pose NDF prior and a gradient-based data augmentation method. In this section, we explain the details of the model and algorithm, as well as the training details.

3.1. Polar-coordinate representation NDF

In this work, we consider the skeleton representation of 2D human poses. As shown in Figure 2, we represent a 2D pose skeleton with J ($J = 15$) vectors, while each vector v_i is represented by the cosine and sine values of its orientation angle ($\theta_i^{(1)}$ and $\theta_i^{(2)}$) and its length (r_i). Thus, the pose

data could be represented as $\mathcal{X} = \{\boldsymbol{x} = (\boldsymbol{v}_1, ..., \boldsymbol{v}_J) | \boldsymbol{v}_i = (\theta_i^{(1)}, \theta_i^{(2)}, r_i) \in \mathbb{R}^3, ||(\theta_i^{(1)}, \theta_i^{(2)})|| = 1 \forall i \in [J]\}$.

The neural distance field for 2D poses is the same as [24], which is $f^{ndf} : \mathcal{X} \to \mathcal{Y} \in \mathbb{R}_{\geq 0}$. The neural distance field f^{ndf} is modeled by a neural network in encoder-decoder structure $f^{ndf} := g^{dec} \circ g^{enc}$. The encoder $g^{enc} : \mathcal{X} \to \mathcal{Z} \in \mathbb{R}^{J \cdot L}$ is a hierarchical structured neural network [1], where L is a hyperparameter and $J \cdot L$ is the dimension of the embedded feature space. The decoder $g^{dec} : \mathcal{Z} \to \mathcal{Y}$ is a simple MLP as in [24].

Training of f^{ndf} is regularized by L2 loss of distance predictions. For a training data $\mathcal{D} = \{(\boldsymbol{x}, d)^{(i)} | i \in [N]\}$ which consists of real pose samples $(\boldsymbol{x}_{gt}, d)^{(i)}$ and fake pose samples $(\boldsymbol{x}_{fk}, d)^{(i)}$ (generated through the process described in 3.2), the loss term for distance prediction is:

$$\mathcal{L}_{real} = \sum_{(\boldsymbol{x}_{gt}, d) \in \mathcal{D}} ||f^{ndf}(\boldsymbol{x}_{gt}) - d||_2,$$
$$\mathcal{L}_{fake} = \sum_{(\boldsymbol{x}_{fk}, d) \in \mathcal{D}} ||f^{ndf}(\boldsymbol{x}_{fk}) - d||_2 \quad (1)$$

In addition, we add a loss term to regularize the gradient of f^{ndf} on real poses as 0 to ensure the NDF model converges better to real poses (represented as $\boldsymbol{x}_{gt}$):

$$\mathcal{L}_{grad} = \sum_{(\boldsymbol{x}_{gt}, d) \in \mathcal{D}} ||\nabla f^{ndf}(\boldsymbol{x}_{gt})||_2. \quad (2)$$

The total loss term is sum of Equation 1 and Equation 2:

$$\mathcal{L}_{total} = \mathcal{L}_{real} + \mathcal{L}_{fake} + \mathcal{L}_{grad} \quad (3)$$

3.2. Fake pose generation

To learn an NDF model, it is necessary to construct a dataset consisting of both real and fake poses, i.e. both samples on and outside the manifold. For real poses, it is straightforward to have zero distance because they are on the manifold. To augment the data, we flipped and interpolated the real poses. We then generate fake poses $\boldsymbol{x}_{fk}$ from erroneous estimates $\hat{\boldsymbol{x}}_{gt}$ from HPE models [7, 23, 27] and the corresponding ground truth pose $\boldsymbol{x}_{gt}$. First, we gather all detection errors into $\mathcal{E} = \{\varepsilon^{(i)} = \hat{\boldsymbol{x}}_{gt}^{(i)} - \boldsymbol{x}_{gt}^{(i)} | i \in [N_{gt}]\}$, where N_{gt} is the number of real poses in the training set. Then, fake poses are generated using the following equation:

$$\boldsymbol{x}_{fk}^{(i)} = u\varepsilon^{(j)} + \boldsymbol{x}_{gt}^{(k)} + \lambda T_{polar}(\boldsymbol{\xi}),$$
$$u \sim U(0, 1), \boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}); i \in [M]; j, k \in [N_{gt}] \quad (4)$$

where M is the number of fake poses to be generated, $\boldsymbol{\xi}$ is random noise in Cartesian coordinates, $T_{polar}(\cdot)$ transforms poses in Cartesian coordinates to polar coordinates, and λ is a small scaling factor. In Equation.4, we introduced a randomly sampled detection noise $\varepsilon^{(j)}$ scaled by u. In this way, we ensure that the model focuses on learning the near-manifold regions that cover the real detection noises.

3.3. Distance definition

The computation of the distance of fake poses follows the procedure similar to the previous work [21, 24], except for the definition of distances. For a given fake pose $\boldsymbol{x}_{fk}$, we first find its K nearest real poses $\{\boldsymbol{x}_{NN}^{(i)} | i \in [K]\}$ using a distance function $dist : (\mathcal{X}, \mathcal{X}) \to \mathcal{Y} \in \mathbb{R}_{\geq 0}$. A common definition of $dist$ is the geodesic distance weighted on each joint:

$$dist(\boldsymbol{x}_1, \boldsymbol{x}_2) =$$
$$\sum_{j=1}^{J} w_j \sqrt{(r_{j,1}\theta_{j,1}^{(1)} - r_{j,2}\theta_{j,2}^{(1)})^2 + (r_{j,1}\theta_{j,1}^{(2)} - r_{j,2}\theta_{j,2}^{(2)})^2}, \quad (5)$$

where w_j^{joint} is the pre-defined weight for joint i, $(\theta_{j,1}^{(1)}, \theta_{j,1}^{(2)}, r_{j,1})$ is the jth element of $\boldsymbol{x}_1$ and $(\theta_{j,2}^{(1)}, \theta_{j,2}^{(2)}, r_{j,2})$ is the jth element of $\boldsymbol{x}_2$. Another possible option for polar coordinate representations is to compute the angular and radial difference separately[4, 11]. Following this idea, we propose an arc-radius-based distance:

$$dist(\boldsymbol{x}_1, \boldsymbol{x}_2) =$$
$$\sum_{j=1}^{J} w_j (r_{j,1} \cdot \arccos(\theta_{j,1}^{(1)}\theta_{j,2}^{(1)} + \theta_{j,1}^{(2)}\theta_{j,2}^{(2)}) + |r_{j,1} - r_{j,2}|). \quad (6)$$

We experimentally show that Equation 6 is a better option than Equation 5 for polar coordinate representations (see Section 4.5). After the K-nearest real poses $\{\boldsymbol{x}_{NN}^{(i)} | i \in [K]\}$ are identified with $dist(\cdot)$, we compute the distance-weighted mean pose as the prior pose of $\tilde{\boldsymbol{x}}$ as:

$$\boldsymbol{x}_{prior} = \sum_{i=1}^{K} w_i^{pose} \boldsymbol{x}_{NN}^{(i)},$$
$$w_i^{pose} =$$
$$(1 - dist(\boldsymbol{x}_{fk}, \boldsymbol{x}_{NN}^{(i)}) / \sum_{k=1}^{K} dist(\boldsymbol{x}_{fk}, \boldsymbol{x}_{NN}^{(k)})) / (K - 1), \quad (7)$$

and re-calculate the distance $dist(\boldsymbol{x}_{fk}, \boldsymbol{x}_{prior})$ as the distance of $\boldsymbol{x}_{fk}$ to the manifold.

3.4. Batch projection augmented training

Training with only supervision from the distance of real/fake poses is not sufficiently effective when the training data are small, so we introduce batch projection-based data augmentation to enhance the training process. We generate a new sample $\boldsymbol{x}_{fk}^{bp}$ from a fake pose $\boldsymbol{x}_{fk}$ by gradient projection, which is similar to [31]:

$$\boldsymbol{x}_{fk}^{bp} = \boldsymbol{x}_{fk} - f^{ndf}(\boldsymbol{x}_{fk}) \cdot \nabla f^{ndf}(\boldsymbol{x}_{fk}). \quad (8)$$

This projection process is performed on the whole batch $B \in \mathcal{X}$ for specific iterations (N^{bp}) during each training

epoch, where each batch consists of both real and fake samples, i.e. $B = \{B_{gt}, B_{fk}\}$. Once a sample is projected close to the manifold (determined with a threshold τ), this sample will be excluded from the batch for the following projections. This iterative batch sample projection process supervises the model to better learn the fields from a fake pose sample to the manifold surface. The training process with batch projection data augmentation is summarized in Algorithm 1.

Algorithm 1 Batch Projection Augmented Training

Require: Training data (X, Y), model $f^{ndf}(\cdot)$, learning rate η, projection iterations N^{bp}, threshold τ
Ensure: Updated model parameters θ
 1: **for** each training epoch **do**
 2:     **for** each batch $B = \{B_{gt}, B_{fk}\} \subset X$ **do**
 3:         **for** $n = 1$ to N^{bp} **do**
 4:             Update $B_{fk} \leftarrow B_{fk} - f^{ndf}(B_{fk}) \cdot \nabla f^{ndf}(B_{fk})$
 5:             **for** each $\boldsymbol{x}_{fk} \in B_{fk}$ **do**
 6:                 **if** $f^{ndf}(\boldsymbol{x}_{fk}) < \tau$ **then**
 7:                     Remove $\boldsymbol{x}_{fk}$ from future projections
 8:                     **break**
 9:                 **end if**
10:             **end for**
11:         **end for**
12:         Compute loss and update model θ
13:     **end for**
14: **end for**
15: **return** Updated parameters θ

3.5. Pose correction process

We adapt the method in [24] to correct 2D poses with a trained NDF pose prior. Given an erroneous 2D pose estimate in Cartesian coordinate $\boldsymbol{x}_{fk}^{cart}$, by setting $\boldsymbol{x}_{fk}^{cart}$ learnable while the parameters of the NDF pose prior fixed, we use the following loss function to correct the pose:

$$\mathcal{L}_{corr} = f^{ndf}(T_{polar}(\boldsymbol{x}_{fk}^{cart})). \tag{9}$$

The correction iteration ends when the output distance of f^{ndf} becomes smaller than a threshold, which is determined using the validation set.

4. Experimental results

In this section, we first introduce the dataset and metric used for evaluating polar coordinate based 2D pose NDF prior (see Section 4.1). Then, we describe the implementation details of the experiments in Section 4.2. Thereafter, we demonstrate the main results of pose correction with the trained prior (Section 4.3), as well as a more detailed inspection into the joint-wise performance (Section

Data set	No. of sequences	No. of poses
	16 (full)	1249
Training	8 (half)	638
	4 (quarter)	280
Validation	4	315
Test	6	441
Total	26	2005

Table 1. Summary of data size and data splits.

4.4). Finally, we performed ablation studies on the proposed method in Section 4.5.

4.1. Dataset and metrics

We evaluated the proposed pose prior using the long jump dataset introduced by Gan et al. [9]. The dataset comprises 26 annotated clips of long jump performances by 22 athletes (11 men and 11 women), captured from broadcast footage of two World Championships. Each clip is recorded at a resolution of 1920×1080 and a frame rate of 25 frames per second. Ground-truth 2D pose annotations and foot–ground contact labels are provided. Using these contact flags, we segmented each clip to retain only the running and take-off phases where pose detection errors are most prevalent. In addition to the ground-truth annotations, we generated pose estimates using three state-of-the-art HPE models: AlphaPose [7], HRNet [23], and ViTPose [27], all applied with corrected bounding boxes detected by YOLOv6 [15] to improve input consistency.

Since the original dataset in [9] does not include a predefined split, we manually defined a custom train–validation–test partition. Among the 26 long jump sequences, 6 sequences were randomly selected for the test set, 4 sequences for the validation set, and the remaining 16 sequences for training. From the training set, a subset of 8 sequences was selected to form the half-training set, and 4 sequences from that subset were further chosen as the quarter-training set. The total number of frames and pose annotations for each split is summarized in Table 1.

We use Percentage of Correct Keypoints (PCK) as the evaluation metric in our experiments. The PCK metric involves a threshold t, which defines the reference distance as t times the distance between the left shoulder and the right hip. A predicted joint is considered correct if its distance to the ground-truth joint is less than or equal to this reference distance. We evaluate performance at thresholds of $t = 0.1$, 0.2, and 0.5. According to [17], a threshold of 0.1 approximately corresponds to a positional tolerance of 6 cm.

4.2. Implementation details

The data were preprocessed to train the NDF pose prior. The experiments are primarily based on pose estimations from ViTPose [27], HRNet [23], and AlphaPose [7]. The

pose skeleton definitions for ViTPose and HRNet follow the COCO WholeBody format [12], while AlphaPose uses the Halpe26 format [7]. These skeletons were converted to a 17-keypoint format, consistent with the definition in [10]. The 2D poses were normalized such that all poses have unit height and their hip joints aligned at the origin.

For the training and validation sequences, we applied horizontal flipping and interpolation with a scaling factor of 5. We then generated 50 times more fake poses than the number of ground-truth poses, using the method described in Section 3.2. These poses were subsequently transformed into either polar coordinates or angular coordinates (as in [21], for ablation studies). The distances between poses were computed using the algorithm detailed in Section 3.3. For each fake pose, we used FAISS [13] to retrieve its 3 nearest-neighbor ground-truth poses for distance calculation.

We trained the NDF model using the Adam optimizer [14] with a learning rate of 0.0001. We assigned equal weights to all loss components in Equation 3. Batch projection was conducted for 20 iterations per epoch across 20 total epochs. For comparison, when batch projection was not used, the model was trained for 400 epochs.

Pose correction on the test set was also performed using the Adam optimizer with a learning rate of 0.0001. During correction, each pose was refined over 100 iterations, or until the predicted distance from the NDF pose prior dropped below a threshold determined on the validation set.

4.3. Overall evaluation

We conducted pose correction experiments using the trained NDF pose prior on the pose detections produced by Alpha-Pose [7], HRNet [23], and ViTPose [27], respectively. The main results are presented in Table 2. The findings indicate that the NDF pose prior substantially enhanced joint detection accuracy.

Evaluation of the raw pose detections reveals that the performances of the three models vary considerably. In general, ViTPose achieves the highest accuracy, while AlphaPose performs the worst under coarser thresholds (PCK@0.2 and PCK@0.5), and HRNet exhibits the lowest accuracy under stricter thresholds (PCK@0.05 and PCK@0.1).

The NDF pose prior consistently improved accuracy across all cases, demonstrating the method's effectiveness and robustness. When comparing models trained on datasets of different sizes, although a larger training set yielded slightly better performance, the improvement margins remained modest. These results suggest that the proposed approach enables the model to learn effectively from small datasets without significant performance degradation. Despite the improved accuracy compared to raw detections, the overall gain is relatively limited. We attribute this pri-

marily to convergence issues in the model, which are further discussed in Section 4.6.

4.4. Joint-wise correction

To inspect the pose correction results in greater detail, we conducted a joint-wise performance analysis, as shown in Figure 3. The mean PCK@0.05, 0.1, and 0.2 values per joint across 50 correction iterations are plotted. Each sub-plot contains 16 curves, each corresponding to a specific joint. At the PCK@0.05 threshold, the most effectively corrected joints include RShoulder (right shoulder), LShoulder (left shoulder), and Nose, while joints such as Head, Thorax, RAnkle (right ankle), and LAnkle (left ankle) exhibit substantial errors that are not well corrected. At the PCK@0.1 level, Head, RAnkle, and LAnkle remain the most error-prone joints, with only the Head showing significant improvement during correction. At the PCK@0.2 threshold, LWrist (left wrist) and RWrist (right wrist) show large initial errors with minimal improvement, whereas RElbow (right elbow) and LElbow (left elbow) demonstrate notable error reduction over iterations. These observations suggest that joints exhibiting less variation in motion (i.e., smaller movement ranges) are corrected more effectively by the NDF pose prior. In contrast, joints with greater motion complexity, such as wrists and ankles, are more challenging to correct. This indicates that additional training data, particularly with a wider range of motion patterns, may help improve the model's ability to refine such joints more accurately.

4.5. Ablation studies

To examine the individual contributions of our proposed components, we conducted a series of ablation studies. Specifically, we focused on four key aspects of our method: (1) the polar coordinate representation, (2) the batch projection-based data augmentation, (3) the arc-radius distance metric, and (4) the gradient loss term defined in Equation 2. To this end, we performed experiments across five different configurations to isolate the impact of each component:

- **Polar baseline**: This baseline setting uses the polar coordinate-based representation for 2D poses. The distances of fake poses are defined using the arc-radius distance, as described in Equation 6. The model is trained using the batch projection-based data augmentation process with both distance-related loss and gradient-related loss.
- **Angular baseline**: This baseline setting follows the angular representation as described in [21]. The distances of fake poses are defined using the angular distance. The model is trained using both batch projection-based data augmentation and gradient-related loss.
- **Polar w/o bp**: This setting is similar to the polar base-

Model	Method	PCK@0.05	PCK@0.1	PCK@0.2	PCK@0.5
AlphaPose	raw	44.88	79.86	91.32	97.11
	full	**45.11**	<u>80.10</u>	**91.62**	**97.13**
	half	<u>45.08</u>	**80.22**	91.56	**97.13**
	quarter	44.94	80.10	<u>91.57</u>	97.12
HRNet	raw	41.18	75.08	94.12	98.05
	full	41.39	<u>76.35</u>	<u>94.32</u>	<u>98.12</u>
	half	<u>41.48</u>	76.20	**94.37**	98.08
	quarter	**41.60**	**76.36**	94.30	**98.13**
ViTPose	raw	51.18	88.62	98.07	99.47
	full	<u>51.78</u>	**90.52**	<u>98.19</u>	<u>99.48</u>
	half	51.35	**90.52**	98.15	<u>99.48</u>
	quarter	**51.95**	90.21	**98.20**	**99.49**

Table 2. Evaluation of the pose correction performance using the trained pose prior. In the 'Method' column, *raw* refers to the uncorrected pose estimations obtained directly from the models listed in the first column. *Full*, *half*, and *quarter* indicate the respective proportions of the training set (as defined in Table 1) used to train the pose prior model.

line, except that batch projection augmentation is not applied during training. To compensate for this, the model is trained for additional epochs until no further improvement is observed on the validation set.

- **Polar w/o grad. loss**: This setting is similar to the polar baseline, except that the gradient loss term is excluded during training.
- **Polar w/o AR. dist.**: This setting is similar to the polar baseline, except that the distances of the fake poses are defined using the geodesic distance (Equation 5).

The ablation study is conducted on pose detections from AlphaPose, using three PCK metrics. The results are summarized in Table 3. The polar baseline achieves the highest accuracy in PCK@0.2, ranks second in PCK@0.1, with only a 0.01% lower accuracy than the best, and is third in PCK@0.05. Therefore, the polar baseline demonstrates the best overall performance.

By comparing the polar baseline with the angular baseline, we observe that the polar representation offers slightly better performance. The results from 'Polar w/o bpj' indicate that the batch sample projection plays a key role in enhancing model performance. The 'Polar w/o grad. loss' yields the lowest performance across all three metrics, suggesting that this loss term significantly influences the learning of a robust NDF prior. Lastly, the 'Polar w/o AR dist.' achieves the second-best accuracy in PCK@0.05 and PCK@0.2, demonstrating that while the geodesic distance performs slightly worse than the proposed arc-radius distance, the difference is not substantial.

4.6. Further studies

We conducted an additional analysis to evaluate the convergence of the pose prior. As shown in Figure 3, the accuracy of some joints decreases over iterations, suggesting that the

Setting	PCK@0.05	PCK@0.1	PCK@0.2
Polar baseline	45.11	<u>80.10</u>	**91.62**
Angular baseline	44.96	**80.11**	91.53
Polar w/o bp	**45.20**	79.99	91.45
Polar w/o grad. loss	44.91	80.02	91.40
Polar w/o AR. dist.	<u>45.15</u>	80.06	<u>91.58</u>

Table 3. Ablation studies on AlphaPose detections

Model	Method	PCK@0.05	PCK@0.1
AlphaPose	baseline	45.11	80.10
	opt. converged	50.31	82.05
HRNet	baseline	41.39	76.35
	opt. converged	46.03	79.33
ViTPose	baseline	51.78	90.52
	opt. converged	57.06	92.34

Table 4. Assessment of the convergence behavior of the NDF pose prior. Baseline performances are compared with the average of the highest accuracy achieved for each pose during the correction process, reported as *opt. converged* in the table.

manifold learned by the NDF pose prior did not fully converge. In this analysis, we compare the baseline performance with the average of the highest accuracy achieved for each pose sample using the same NDF pose prior. The results presented in Table 4 demonstrate that while the model achieves significantly improved pose corrections during the refinement process, it does not consistently converge to the optimal correction. This indicates that even higher performance could be attained if the convergence issue were effectively addressed.
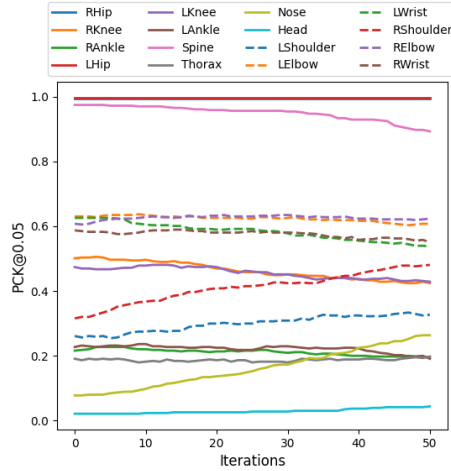
(a) PCK@0.05
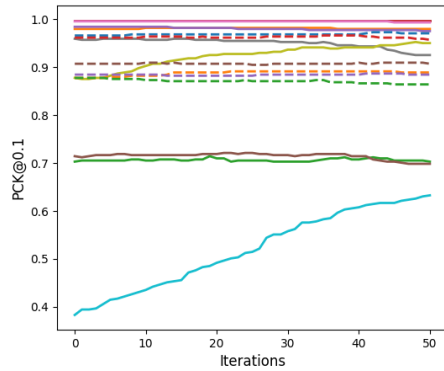
(b) PCK@0.1
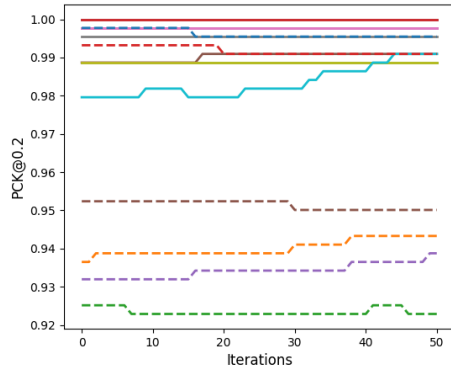
(c) PCK@0.2

Figure 3. Mean PCK per joint over 50 iterations.

5. Discussions and conclusions

In this work, we proposed a polar coordinate-based representation for learning 2D pose prior with the neural distance field. To facilitate the training of the pose prior, we also pro-

posed gradient loss term and an arc-radius distance for polar coordinate representations. To enable model training with small datasets, we proposed a data augmentation method to generate new sample batches by gradient-based projection. The experimental results showed a consistency improvement of the pose key point accuracy, regardless of the source model for the pose estimations. We also show that the model could be trained with less than a quarter of the original training data (280 poses out of 1249 poses) without a big sacrifice of performance.

Despite the effectiveness and robustness of the proposed method, there are still two major issues. The first is that the model tends to be less capable of correcting the large motion joints (wrists and ankles), which could be caused by insufficiency of data varieties. Moreover, currently, the model still has difficulty with convergence, which greatly hinders the performance. This article shows that the model has potential to achieve a much better result if the convergence issue were properly resolved.

References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7144–7153, 2019. 2, 4

[2] Sian Barris and Chris Button. A review of vision-based motion analysis in sport. *Sports medicine*, 38:1025–1043, 2008. 1

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[4] Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Chang Huang, and Wenyu Liu. Polar parametrization for vision-based surround-view 3d detection. *arXiv preprint arXiv:2206.10965*, 2022. 4

[5] Julian Chibane, Gerard Pons-Moll, et al. Neural unsigned distance fields for implicit function learning. *Advances in Neural Information Processing Systems*, 33:21638–21652, 2020. 2, 3

[6] Bhat Dittakavi, Divyagna Bavikadi, Sai Vikas Desai, Soumi Chakraborty, Nishant Reddy, Vineeth N Balasubramanian, Bharathi Callepalli, and Ayon Sharma. Pose tutor: an explainable system for pose correction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3540–3549, 2022. 3

[7] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7157–7173, 2022. 1, 4, 5, 6

[8] Mykyta Fastovets, Jean-Yves Guillemaut, and Adrian Hilton. Athlete pose estimation from monocular tv sports footage. In *Proceedings of the IEEE Conference on Com-

puter Vision and Pattern Recognition Workshops, pages 1048–1054, 2013. 3

[9] Qi Gan, Mounîm A El-Yacoubi, Eric Fenaux, Stéphan Clémençon, et al. Human pose estimation based biomechanical feature extraction for long jumps. In *2024 16th International Conference on Human System Interaction (HSI)*, pages 1–6. IEEE, 2024. 1, 2, 5

[10] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 6

[11] Ran Iwamoto, Ryosuke Kohita, and Akifumi Wachi. Polar embedding. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 470–480, 2021. 4

[12] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision*, pages 196–214. Springer, 2020. 6

[13] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 6

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[15] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, Yiduo Li, Bo Zhang, Yufei Liang, Linyuan Zhou, Xiaoming Xu, Xiangxiang Chu, Xiaoming Wei, and Xiaolin Wei. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications, 2022. 5

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 1

[17] Katja Ludwig, Sebastian Scherer, Moritz Einfalt, and Rainer Lienhart. Self-supervised learning for human pose estimation in sports. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2021. 1, 3, 5

[18] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[19] Mahdiar Nekoui, Fidel Omar Tito Cruz, and Li Cheng. Falcons: Fast learner-grader for contorted poses in sports. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 900–901, 2020. 1

[20] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of*

[21] Dripta S Raychaudhuri, Calvin-Khang Ta, Arindam Dutta, Rohit Lal, and Amit K Roy-Chowdhury. Prior-guided source-free domain adaptation for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14996–15006, 2023. 1, 2, 3, 4, 6

[22] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11488–11499, 2021. 3

[23] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 4, 5, 6

[24] Garvita Tiwari, Dimitrije Antić, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision*, pages 572–589. Springer, 2022. 2, 3, 4, 5

[25] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 2

[26] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34:14955–14966, 2021. 2, 3

[27] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 1, 4, 5, 6

[28] Calvin Yeung, Kenjiro Ide, and Keisuke Fujii. Autosoccerpose: Automated 3d posture analysis of soccer shot movements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3214–3224, 2024. 1

[29] Dan Zecha, Moritz Einfalt, Christian Eggert, and Rainer Lienhart. Kinematic pose rectification for performance analysis and retrieval in sports. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1791–1799, 2018. 2, 3

[30] Dan Zecha, Moritz Einfalt, and Rainer Lienhart. Refining joint locations for human pose tracking in sports videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2, 3

[31] Junsheng Zhou, Baorui Ma, Shujuan Li, Yu-Shen Liu, and Zhizhong Han. Learning a more continuous zero level set in unsigned distance fields through level set projection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3181–3192, 2023. 4