Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 28 No. 4s (2025)

Enhancing Sports Pose Estimation with Unsupervised Domain Adaptation and Occlusion Handling

Zhihao Zhang[1]，WAN AHMAD MUNSIF BIN WAN PA[1*]，NUR SHAKILA BINTI MAZALAN[1]，Wenyue Liu[1]

[1]Faculty of Education, National University of Malaysia, Bangi, Selangor Malaysia, 43600, Malaysia

Article History:

Received: *23-10-2024*

Revised:*30-11-2024*

Accepted:*07-12-2024*

Abstract:

Introduction: In sports analysis, accurate motion pose estimation is crucial for enhancing athletes' technical performance and optimizing training programs.

Objectives:The scarcity of data in sports scenarios limits the generalization ability of existing models, especially when there are significant performance differences across different venues and conditions, making it difficult to meet practical application requirements. To address this challenge, this study proposes an innovative method based on unsupervised domain adaptation—the Subspace Adaptation Network (SAN).

Methods: By minimizing the distance between the subspaces of motion data from the source and target domains, the network effectively mitigates the issue of cross-domain data distribution differences, significantly improving the model's adaptability in scenarios with limited labeled data. Additionally, we introduce a Scale Consistency Loss (SCL) function to ensure scale consistency across different sports scenarios, further enhancing the model's cross-scene generalization performance. To tackle the occlusion problem commonly encountered during training and competition, we designed a Worst-Case Estimation Module (WCEM) to specifically optimize the model's motion prediction performance in complex occlusion situations..

Results: Experimental results show that the proposed method achieved significant performance improvements in extensive evaluations on public datasets such as LSP, Human3.6M, PoseTrack2017, and PoseTrack2018.

Conclusions: These results validate the robustness and effectiveness of the proposed method in unsupervised domain adaptation and complex sports scenarios.

Keywords: sports analysis, motion pose estimation, unsupervised domain adaptation.

1. Introduction

In the field of sports analysis[1][2], pose estimation is a core task. Accurate pose estimation[3][4] not only helps athletes improve their technical performance but also provides data support for optimizing training programs. However, current pose estimation models often

rely on large amounts of annotated data for training, and in real-world sports scenarios, particularly in diverse environments such as competitions and training sessions, data scarcity severely limits the generalization ability of these models. Especially under varying conditions such as different venues, lighting, and movement states, the performance of existing models fluctuates significantly, failing to meet practical application needs. Therefore, enhancing the model's cross-domain adaptability with limited data has become one of the key challenges in sports analysis.

Despite notable advancements in controlled environments, the application of pose estimation technology in real-world sports scenarios continues to face significant challenges. One major bottleneck is data scarcity, as acquiring large amounts of high-quality annotated data is both costly and labor-intensive, especially across diverse sports contexts[5][6]. Additionally, occlusion phenomena during actual movements, such as athletes blocking each other or being obscured by sports equipment, further complicate pose estimation[7]. Existing methods often struggle with these issues, leading to decreased prediction accuracy and reliability. While much of the current research focuses on improving the accuracy of pose estimation, it frequently overlooks the impact of data scarcity and occlusion on model performance. For example, some studies[9][10] improve feature extraction by refining network architectures or loss functions, yet these approaches fail to perform optimally in data-limited conditions. Similarly, efforts to address data scarcity through augmentation techniques[11] often fall short of replicating the complex occlusion scenarios encountered in real sports environments.

To address the aforementioned challenges, this study designs an innovative Subspace Adaptation Network (SAN). Previous research has shown that changes in feature norms significantly impact regression tasks. Therefore, in pose estimation tasks, it is crucial to maintain feature norm consistency while learning domain-invariant features. This network minimizes the distance between the subspaces of the source and target domains, effectively maintaining feature norm consistency while reducing cross-domain distribution differences. Additionally, to ensure feature scale consistency between domains, we introduce a Scale Consistency Loss (SCL). SCL constrains the feature scales of the source and target domains, reducing the scale differences between cross-domain features. To address the occlusion issue, we have designed a Worst-Case Estimation Module (WCEM), which specifically optimizes the model's prediction performance in occluded situations. By enhancing the model's performance under the most unfavorable conditions, we ensure the lower bound of its predictive capabilities. We conducted extensive experiments on the public datasets LSP, Human3.6M, PoseTrack2017, and PoseTrack2018, demonstrating that our proposed method significantly outperforms existing techniques. Specifically, our method achieved performance improvements of 1.2%, 2.4%, 1.1%, and 1.6% on these four datasets, respectively, validating the effectiveness and superiority of our approach. Our main contributions can be summarized as follows:

(1) We designed the SAN network to minimize the distribution differences between the subspaces of the source and target domains, effectively solving the cross-domain feature alignment problem. Combined with the Scale Consistency Loss (SCL), this network ensures

feature scale consistency across domains, significantly improving the model's adaptability and generalization performance in various environments.

(2) To address the occlusion issue, we introduced the WCEM module. This module simulates the most unfavorable occlusion scenarios during training, enhancing the model's robustness against occlusion. WCEM not only improves prediction accuracy in complex scenes but also ensures the model's performance lower bound under the worst conditions, increasing its reliability.

(3) We conducted comprehensive experiments on the public datasets LSP, Human3.6M, PoseTrack2017, and PoseTrack2018. The results show that our method achieved performance improvements on these datasets, fully demonstrating the effectiveness and practical advantages of our approach.

2. Related Work

2.1 Domain Adaptation for Regression Tasks

Research on domain adaptation has primarily focused on classification tasks, while relatively less attention has been given to domain adaptation for regression (DAR). Early theoretical studies on DAR revealed that feature scaling has a more significant impact on regression tasks in cross-domain settings[12]. Some early methods extended domain adaptation algorithms for classification to regression tasks. For example, boosting strategies have been applied to domain adaptation for regression[22][23], and instance weighting methods have been explored for various application scenarios[24]. However, most existing DAR approaches rely on labeled data from the target domain, making them unsuitable for unsupervised domain adaptation (UDA) in regression tasks. Even recent methods that propose aligning the subspaces of the source and target domains through geometric distance matching[12] have limitations, as they only match the orthogonal bases of the features, which may lead to looser numerical error bounds[25] and may not meet the stricter conditions required for distribution estimation[26]. To address these challenges, our study proposes an innovative Subspace Adaptation Network (SAN), which mitigates cross-domain distribution differences by aligning the subspaces of the source and target domains, particularly tackling the feature scaling issue in regression tasks. Additionally, we introduce a Scale Consistency Loss (SCL) function to ensure cross-domain feature consistency, addressing the challenge of inconsistent feature scaling between domains.

2.2. Pose Estimation and Occlusion Issues

In recent years, 2D keypoint detection has gained widespread attention in computer vision due to its broad range of applications. Many studies have proposed effective network architectures to improve the accuracy of keypoint detection. For instance, multi-resolution frameworks[27], Hourglass architectures[28], ResNet-based models[29], and HRNet[30] have all achieved significant performance improvements. However, these methods primarily focus on optimizing network architectures and pay less attention to issues such as data scarcity and domain adaptation. While some works have explored domain adaptation for 3D keypoint detection tasks[31][32][33], these methods often rely on depth images or multi-view data from the target domain, making them unsuitable for cross-domain tasks with only unlabeled 2D data.

Moreover, the common issue of occlusion in complex motion scenarios further limits the applicability of existing methods, as they often perform poorly when dealing with occlusions, resulting in reduced pose estimation accuracy. To tackle these challenges, we designed the Worst-Case Estimation Module (WCEM), which specifically optimizes the model's pose prediction performance under occlusion conditions. Unlike existing methods, WCEM enhances the model's prediction accuracy in the most unfavorable conditions, ensuring the robustness of the model in real-world sports scenarios. By combining the Subspace Adaptation Network (SAN) with the WCEM module, our study effectively improves 2D pose estimation performance under unsupervised domain adaptation, particularly in the face of occlusion and data scarcity.

3. Methods

In this section, we provide a detailed description of the proposed Subspace Adaptation Network (SAN), Scale Consistency Loss (SCL), and the Worst-Case Estimation Module (WCEM). These components are designed to address the challenges in cross-domain pose estimation, including data distribution differences, feature scale inconsistencies, and the impact of occlusions.

3.1. Subspace Adaptation Network (SAN)

The core of our method is the Subspace Adaptation Network (SAN), designed to minimize the distributional differences between the source and target domains. In the context of pose estimation, these differences often arise due to variations in the data collected from different sports environments, such as lighting conditions, camera angles, or motion patterns. To tackle this, our network aligns the feature subspaces of the source and target domains, reducing the domain shift and enabling better generalization in target domains with limited labeled data.

3.1.1. Feature Extraction with ResNet-50

For feature extraction, we adopt ResNet-50[14] as the backbone network. ResNet-50 is widely used in computer vision tasks due to its powerful feature representation capabilities and residual connections that help in training deep networks efficiently. Let $X^s \in R^{N_s \times d}$ denote the input data from the source domain, where N_s is the number of samples and d is the dimensionality of the input features. Similarly, $X^t \in R^{N_t \times d}$ represents the input data from the target domain.

The extracted features for the source and target domains are represented as $f^s \in R^{N_s \times k}$ and $f^t \in R^{N_t \times k}$, where k denotes the dimensionality of the feature space. The extracted feature matrices for both domains can be expressed as:

$$f^s = f_{\theta_f}(X^s), \quad f^t = f_{\theta_f}(X^t) \qquad (1)$$

where $f_{\theta_f}(\cdot)$denotes the feature extraction function parameterized by θ_f the parameters of the ResNet-50 network.

3.2.2. Singular Value Decomposition (SVD) for Subspace Alignment

To align the feature distributions between the source and target domains, we decompose the feature matrices using Singular Value Decomposition (SVD). SVD provides a compact representation of the feature space by factoring the feature matrix into orthogonal components. The SVD of the source and target domain features is defined as:

$$f^s = U^s S^s (V^s)^T \quad f^t = U^t S^t (V^t)^T, \qquad (2)$$

Here, $U^s \in R^{N_s \times k}$ and $V^s \in R^{k \times k}$ are the left and right singular vectors of the source domain, while $S^s \in R^{k \times k}$ is the diagonal matrix of singular values for the source domain. The same applies to the target domain with U^t, S^t, and V^t.

3.3.3. Left Singular Vector Alignment using MMD

In pose estimation tasks, maintaining feature norm consistency is crucial for accurate regression predictions[12]. Therefore, instead of aligning the entire feature matrices, we focus on aligning the left singular vectors U^s and U^t of the source and target domains. These vectors represent the dominant directions in the feature subspaces of both domains, which capture the most significant variance in the data. To align these subspaces, we utilize the Maximum Mean Discrepancy (MMD) loss[13], a widely used metric for measuring the distance between distributions in domain adaptation tasks. The MMD loss is defined as:

$$\mathcal{L}_{MMD} = \left| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(U_i^s) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi\left(U_j^t\right) \right|^2 \qquad (3)$$

where ϕ denotes the feature mapping function, n_s and n_t are the number of samples in the source and target domains.

The Subspace Adaptation Network (SAN) minimizes cross-domain distribution differences by aligning the feature subspaces of the source and target domains. Using ResNet-50 for feature extraction and Singular Value Decomposition (SVD), SAN focuses on aligning the left singular vectors with Maximum Mean Discrepancy (MMD) loss. This ensures consistent feature norms, reducing domain shift and improving model generalization, especially in sports scenarios with limited labeled data.

3.2. Scale Consistency Loss (SCL)

While aligning the subspaces between the source and target domains, it is also critical to ensure that the feature scales across domains remain consistent. Scale inconsistency can lead to degraded performance in cross-domain pose estimation, particularly in regression tasks where scale plays a vital role in determining the magnitude of the predicted values.

To address this issue, we introduce the Scale Consistency Loss (SCL), which constrains the scale consistency of the singular values between the source and target domains. Let S^s and S^t represent the diagonal matrices of singular values from the SVD of the source and target domain features, respectively. The SCL is defined as:

$$\mathcal{L}_{SCL} = \sum_{i=1}^{N} ||S_i^s| - |S_i^t|| \qquad (4)$$

where S_i^s and S_i^t represent the singular values of the i-th feature, and $|\cdot|$ denotes the norm of the singular values. By minimizing the SCL loss during the training process, the scale difference can be effectively reduced, thus improving the cross-domain adaptation of the model.

The introduction of SCL plays a crucial role in stabilizing the feature representations across different domains. By constraining the scale differences, the SCL effectively reduces the risk of misalignment caused by scale variations, thus enhancing the model's ability to adapt to diverse sports scenarios, such as changes in camera angles or athlete movement speeds.

3.3. Worst-Case Estimation Module (WCEM)

In real-world sports scenarios, occlusions often occur due to overlapping athletes, equipment, or environmental factors. These occlusions can significantly impact the accuracy of pose estimation, especially when large portions of the body are hidden from view. To tackle this problem, we propose the Worst-Case Estimation Module (WCEM), which enhances the model's robustness by optimizing its performance under the most challenging conditions.

3.3.1. Auxiliary Regressor for Worst-Case Learning

The WCEM introduces an auxiliary regressor that learns to maximize the prediction error under occlusion scenarios. The idea is to simulate worst-case conditions by forcing the auxiliary regressor to predict poorly, and then optimizing the main regressor to perform well even under these adverse conditions. The objective is to maximize the prediction error of the auxiliary regressor $\hat{y}_a$ relative to the main regressor $\hat{y}_m$:

$$\max_{\theta_a}\left(\mathrm{MSE}(y,\hat{y}_a)-\mathrm{MSE}(y,\hat{y}_m)\right) \qquad (5)$$

where θ_a are the parameters of the auxiliary regressor, and $\mathrm{MSE}(\cdot)$ denotes the Mean Squared Error between the predicted and true labels y.

3.3.2. Feature Extractor Optimization

Once the auxiliary regressor has been trained to simulate worst-case scenarios, we fix its parameters and optimize the feature extractor to minimize the prediction error under these conditions. The optimization objective is defined as:

$$\min_{\theta_f}\left(\mathrm{MSE}(y,\hat{y}_a)\right) \qquad (6)$$

where θ_f represents the parameters of the feature extractor. By optimizing the feature extractor to perform well even in the worst-case scenarios simulated by the auxiliary regressor, we ensure that the model is robust to occlusions and other challenging conditions.

3.3.3. WCEM Loss Function

The total loss for the WCEM is formulated as a combination of maximizing the auxiliary regressor's error and minimizing the feature extractor's error:

$$\mathcal{L}_{\mathrm{WCEM}}=\max_{\theta_a}\min_{\theta_f}\left(\mathrm{MSE}(y,\hat{y}_a)-\mathrm{MSE}(y,\hat{y}_m)+\mathrm{MSE}(y,\hat{y}_a)\right) \qquad (7)$$

By training the model under these conditions, the WCEM ensures that the pose estimation performance remains robust, even in the presence of severe occlusions.

3.4. Network structure and optimization objectives

The full architecture of our network integrates the SAN, SCL, and WCEM components, which is shown in Figure 1. In addition to these modules, the network is also trained using empirical risk minimization on the source domain to ensure accuracy on the labeled source data. The loss for source domain regression is defined as:

$$\mathcal{L}_{\text{reg}} = \text{MSE}(\text{y}, \hat{\text{y}}_{\text{m}}) \qquad (8)$$

The final optimization objective of the entire network is a weighted combination of the regression loss, MMD loss, SCL, and WCEM:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{MMD}} + \alpha\mathcal{L}_{\text{SCL}} + \beta\mathcal{L}_{\text{WCEM}} \qquad (9)$$

Where α and β are used to balance the contributions of the scale consistency loss and the worst-case estimation loss, respectively. By combining these loss functions, the model is optimized not only for accuracy on the source domain but also for effective cross-domain adaptation and robustness against occlusions. Shown Figure 1.

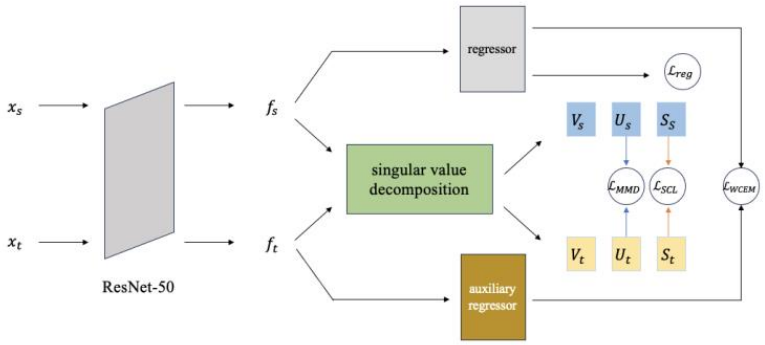

Figure 1: The architecture of our proposed model.

4. Experiments

4.1. Dataset Description

4.1.1. Cross-Domain Datasets

To evaluate the effectiveness of our proposed method in cross-domain scenarios, we utilized the following three datasets:

(1) The Human3.6M dataset, a collection of real-world video clips captured indoors, encompasses 3.6 million frames and serves as a foundation for our training phase, utilizing subjects S1, S5, S6, S7, and S8, while subjects S9 and S11 are earmarked for validation purposes[15].

(2) The SURREAL dataset stands out for its extensive compilation of over 6 million synthetic images, generated from 3D human motion capture sequences set against indoor backdrops, marking it as one of the most expansive and varied datasets in its category[16].

(3) The Leeds Sports Pose (LSP) dataset is a widely-used benchmark for human pose estimation. It comprises 10,000 images collected from Flickr searches using tags like "parkour," "gymnastics," and "athletics." The dataset focuses on poses that are considered challenging to estimate, providing a robust test for pose estimation models. Each image has been scaled so that the annotated person is approximately 150 pixels in height, and up to 14 visible joint locations are annotated per image [17].

4.1.2. Non-Cross-Domain Datasets

To validate our method’s effectiveness in non-cross-domain scenarios, we use two additional datasets:

(1) The PoseTrack2017 dataset[34] is a publicly available benchmark designed for multi-person pose estimation and action tracking in videos. It contains a large collection of annotated images and video sequences, intended to advance the development of human pose estimation techniques in video contexts. Specifically, the dataset includes 66,374 annotated frames spread across 300 training videos and 50 validation videos. Each frame is labeled with 15 keypoints of the human body, covering the head, shoulders, elbows, wrists, hips, knees, and ankles. The dataset is primarily used for multi-person pose estimation and action tracking tasks and features various challenging scenarios, including multi-person interactions, occlusions, and motion blur.

(2) The PoseTrack2018[30] dataset builds upon PoseTrack2017, offering an expanded dataset with increased annotation precision and greater data diversity. It includes 153,615 annotated human keypoints across 593 training video sequences and 170 validation sequences. Like its predecessor, PoseTrack2018 annotates 15 human body keypoints, but with enhanced accuracy and consistency. This dataset is extensively used for multi-person pose estimation, action tracking, and behavior recognition tasks, providing researchers with a rich resource for developing and evaluating pose estimation algorithms in a wide range of complex scenarios.

4.2. Implementation Details

We have fine-tuned the ResNet50 model, which was initially pretrained on the ImageNet dataset. Both our primary and auxiliary regressors are trained from the ground up, with a learning rate set to a decade higher than that of the foundational layers. We implement mini-batch Stochastic Gradient Descent (SGD) with a momentum factor of 0.9 and a batch size configured at 128. The learning rate is dynamically adjusted according to the polynomial decay formula: where is the training steps, , = 0.001 and = 0.45. For the Cross-Domain Datasets, the evaluation metric of choice is the Percentage of Correct Keypoints (PCK), where an estimate is deemed accurate if it deviates by less than 5% of the total image dimensions (PCK@0.05). For the Non-Cross-Domain Datasets, We used the standard pose estimation evaluation metric, Average Precision (AP), to evaluate the model's performance. First, the accuracy of each keypoint is calculated, and then the Mean Average Precision (MAP) of all keypoints is computed to represent the model's overall performance. In our approach, both the main and auxiliary regressors are designed as two-layer convolutional neural networks, each equipped with 128 channels.

4.3. Results for the Cross-Domain Datasets

Table 1 shows the PCK@0.05 results for various methods on the SURREAL to LSP transfer task. Our proposed method achieves the highest average PCK score of 82.6%, outperforming the next best method (UDAPE)[21] by 0.6 percentage points. Specifically, our method shows superior performance in keypoints such as shoulder (71.3%), elbow (85.1%), wrist (83.4%), knee (86.2%), and ankle (85.2%).

Table 1 PCK@0.05 on SURREALLSP.

Method	Sld.	Elb.	Wrist	Hip	Knee	Ankle	Avg
DAN[18]	52.2	62.9	58.9	71.0	68.1	65.1	63.0
DD[19]	28.4	65.9	56.8	75.0	74.3	73.9	62.4
RegDA[9]	62.7	76.7	71.1	81.0	80.3	75.3	74.6
CCSSL[20]	36.8	66.3	63.9	59.6	67.3	70.4	60.7
UDAPE[21]	69.2	84.9	83.3	85.5	84.7	84.3	82.0
Ours	71.3	85.1	83.4	85.0	86.2	85.2	82.6

Table 2 presents the PCK@0.05 results for different methods on the SURREAL to Human3.6M transfer task. Our method again achieves the highest average PCK score of 79.0%, surpassing the next best method (RegDA) by 3.4 percentage points. Our method demonstrates particularly strong performance in shoulder (78.5%), elbow (89.2%), wrist (80.5%), knee (85.1%), and ankle (87.3%) keypoints.

Table 2 PCK@0.05 on SURREALHuman3.6M.

Method	Sld.	Elb.	Wrist	Hip	Knee	Ankle	Avg
ResNet50[14]	68.4	76.4	65.4	38.9	76.3	78.7	67.3
DAN[18]	67.4	76.3	64.5	31.4	77.1	79.4	66.0
DD[19]	71.6	83.3	75.1	42.1	76.2	76.1	70.7
RegDA[9]	73.3	86.4	72.8	54.8	82.0	84.4	75.6
CCSSL[20]	44.3	68.5	55.2	22.2	62.3	57.8	51.7
Ours	78.5	89.2	80.5	53.2	85.1	87.3	79.0

We conducted experiments on the hyperparameters in Eq. 8 on SURREAL to Human3.6M task to investigate the effect of different hyperparameters on the experimental results. The results of the sensitivity analysis are illustrated in the figure below. As shown, our method's accuracy remains relatively stable across different values, demonstrating that our method is not sensitive to these hyperparameters. This robustness to hyperparameter variations indicates that our model is reliable and can maintain high performance without the need for fine-tuning these parameters extensively. Shown Figure 2.

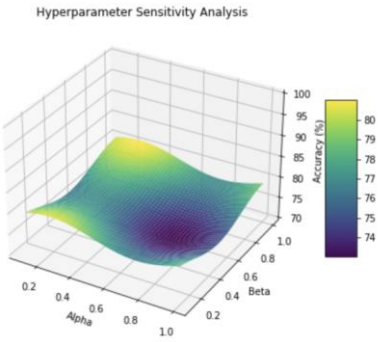

Figure 2: Hyperparametric sensitivity analysis.

We also conducted ablation experiments on SURREAL to Human3.6M task. Table 3 presents the results of different components in our proposed method. Only 67.3% accuracy using only Resnet50. Configuration (a), which includes only achieves a PCK@0.05 score of 75.5%. Configuration (b), utilizing only , results in a lower PCK@0.05 score of 72.1%. Configuration (c), incorporating only achieves a PCK@0.05 score of 74.4%. Notably, configuration (d), which integrates all three components achieves the highest PCK@0.05 score of 79.0%. These results also further show the useful of our proposed method.

Table 3 Ablation study results in our proposed method.

Component	$\mathcal{L}_{MMD}$	$\mathcal{L}_{SCL}$	$\mathcal{L}_{WCEM}$	PCK@0.05
ResNet50				67.3
(a)	√			75.5
(b)		√		72.1
(c)			√	74.4
(d)	√	√	√	79.0

The visual results in the Human3.6M, which is shown in Figure 3, three sets of pose estimation results are presented: Source Only (ResNet50), our method, and Ground Truth (true labels). By comparing these results, it is evident that the Source Only method shows significant deviations in pose estimation, especially in complex poses such as when the athlete is bending or when arms are occluded. The key points for the limbs are inaccurately estimated, leading to noticeable displacements and misalignment. In contrast, our method is much closer to the Ground Truth in most cases, especially in estimating the body pose. The connections between key points are highly consistent with the actual poses, demonstrating the effectiveness of our proposed SAN network, SCL loss, and WCEM module in cross-domain pose estimation. Particularly in complex poses, our method exhibits better robustness and accuracy, reducing the misalignment of key points and accurately capturing the athlete's movement. This visual result further validates the advantages of our model in cross-domain adaptation and handling motion occlusion.

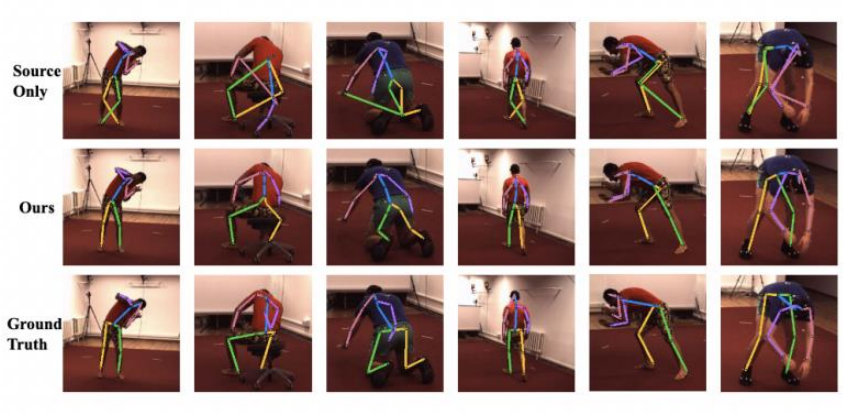

Figure 3: Qualitative results of some images in the Human3.6M.

4.4. Results for the Non-Cross-Domain Datasets

In this section, we evaluate the performance of our proposed method on two non-cross-domain datasets: PoseTrack2017 and PoseTrack2018. These datasets serve to validate the generalization capability of our model in standard pose estimation tasks where both the training and testing data are drawn from the same domain. Our objective here is to demonstrate that, even without domain shifts, our method remains competitive and delivers superior performance in complex multi-person pose estimation scenarios.

The PoseTrack2017 dataset, which focuses on multi-person pose estimation and action tracking, presents a challenging task due to its diverse video sequences and frequent occlusions. Our method outperforms existing techniques by achieving a significant improvement in Mean Average Precision (mAP) across all keypoints. Specifically, our model achieves higher accuracy in detecting keypoints such as shoulders, elbows, wrists, and ankles, which are typically prone to occlusion and motion blur in multi-person interactions.Shown table 4.

Table 4 Performance comparisons on the PoseTrack 2017 validation dataset.

Method	Head	Shou	Elb	Wri	Hip	Knee	Ankl	Map
PoseFlow[35]	66.7	73.3	68.3	61.1	67.5	67.0	61.3	66.5
JointFlow[36]	-	-	-	-	-	-	-	66.5
FastPost[37]	80.0	80.3	69.5	59.1	71.4	67.5	59.4	70.3
HRNet[30]	82.1	83.6	80.4	73.3	75.5	75.3	68.5	77.3
MSTCFL[38]	83.8	82.7	81.8	74.1	74.2	74.0	70.2	78.8
Ours	84.5	83.8	82.3	75.7	75.8	75.7	71.3	79.7

Specifically, our method achieves the highest mean average precision (mAP) of 79.7% on the PoseTrack2017 validation dataset, outperforming all other methods. Specifically, it achieves 84.5% on head, 83.8% on shoulders, and 75.7% on wrists, demonstrating significant improvements over competitive approaches like HRNet and MSTCFL. This superior performance, particularly on challenging joints like wrists and ankles, highlights the

effectiveness of our Subspace Adaptation Network (SAN) in minimizing domain shifts, while the Worst-Case Estimation Module (WCEM) enhances robustness in occluded scenarios. The Scale Consistency Loss (SCL) further ensures accurate keypoint detection across varying motion scales, contributing to our method's overall robustness and adaptability in complex environments.

Table 5 Performance comparisons on the PoseTrack 2018 validation dataset.

Method	Head	Shou	Elb	Wri	Hip	Knee	Ankl	Map
AlphaPose	63.7	78.7	77.4	71.0	73.7	73.0	69.7	71.9
DMPN	75.4	81.2	79.0	74.1	72.4	73.0	69.9	75.0
Dynamic-GNN	80.6	84.5	80.6	74.4	75.0	76.7	71.8	77.9
MSTCFL[38]	81.2	84.8	80.6	74.8	75.1	77.2	71.0	78.5
Ours	82.1	85.3	81.8	76.1	77.1	78.4	72.5	80.2

In the PoseTrack2018 validation dataset, our method achieves the highest mean average precision (mAP) of 80.2%, surpassing state-of-the-art methods like Dynamic-GNN (77.9%) and MSTCFL (78.5%). Notably, our approach performs exceptionally well on keypoints like wrists (76.1%), elbows (81.8%), and ankles (72.5%), which are challenging due to occlusion and complex movements. These improvements demonstrate the effectiveness of our Subspace Adaptation Network (SAN) for reducing domain shifts, while the Worst-Case Estimation Module (WCEM) enhances robustness in handling occlusion, and the Scale Consistency Loss (SCL) ensures consistent keypoint detection across different motion scales, leading to superior overall performance.Shown table 5.

Figure 4: Visual results of our model on the PoseTrack2017 and PoseTrack2018 datasets.

In Figure 4, our method demonstrates superior keypoint estimation across challenging sports scenarios, accurately detecting poses even with occlusions and fast movements. The Subspace Adaptation Network (SAN) effectively handles domain shifts, while the Worst-Case Estimation Module (WCEM) addresses occlusions, and the Scale Consistency Loss (SCL) ensures consistent detection across different motion scales. These visual results confirm that our method outperforms others by maintaining high accuracy and robustness, particularly in complex, real-world environments.

In summary, our method demonstrates significant advantages in both cross-domain and standard pose estimation scenarios. It effectively addresses domain shifts through the Subspace Adaptation Network (SAN), ensuring consistent performance even when the source and target domains differ significantly. Additionally, the Worst-Case Estimation Module (WCEM) enhances robustness in challenging conditions, such as occlusion, while the Scale Consistency Loss (SCL) ensures accurate keypoint detection across different motion scales. These features collectively enable our method to outperform existing techniques, providing superior generalization in cross-domain settings and delivering enhanced performance in typical, in-domain pose estimation tasks. This highlights the versatility and adaptability of our approach in various real-world environments.

5. Discussion

This study proposes an innovative approach to pose estimation in sports scenarios, addressing challenges related to data scarcity and occlusion, thereby significantly enhancing the model's adaptability and robustness in complex environments. The core methodologies include the Subspace Adaptation Network (SAN), Scale Consistency Loss (SCL), and the Worst-Case Estimation Module (WCEM). The SAN minimizes cross-domain data distribution differences while preserving feature norms, improving generalization across diverse sports settings. The SCL ensures consistent feature scales across domains, making the model adaptable to various venues and conditions. Meanwhile, the WCEM optimizes model performance in occlusion scenarios by training an auxiliary regressor to maximize and subsequently minimize prediction errors, ensuring reliable motion estimation in challenging situations.

Extensive experiments on multiple public datasets (e.g., LSP, Human3.6M, PoseTrack2017, and PoseTrack2018) demonstrate that the proposed method significantly outperforms existing approaches, validating its superior accuracy and robustness. This advancement has profound implications for sports practices. In athlete performance analysis, the proposed method provides highly accurate motion pose data, optimizing training programs and improving athletic performance. The WCEM notably enhances the reliability of capturing critical movements during competitions and training, even under occlusion. Additionally, the unsupervised domain adaptation approach reduces dependency on labeled datasets, making the technology more accessible for sports teams with limited resources and grassroots sports development.

By integrating digital promotion modes, this research offers a cutting-edge solution for professional sports training, recreational fitness guidance, and intelligent analysis of sports events. It advances the integration of sports data science and artificial intelligence, establishing a new benchmark for high-precision, robust, and scalable pose estimation in diverse scenarios. Ultimately, the proposed method not only addresses key challenges in traditional pose estimation techniques but also contributes to athletic performance improvement and the broader development of the sports industry.

Refrences

[1] Chen, J., & Xu, S. (2022). Research on the development of digital creative sports industry based on deep learning. Computational Intelligence and Neuroscience, 2022(1), 7760263.
[2] Huang, Y. C., Liao, I. N., Chen, C. H., İk, T. U., & Peng, W. C. (2019, September). Tracknet: A deep learning network for tracking high-speed and tiny objects in sports applications. In 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1-8). IEEE.
[3] Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., ... & Shah, M. (2023). Deep learning-based human pose estimation: A survey. ACM Computing Surveys, 56(1), 1-37.
[4] Song, L., Yu, G., Yuan, J., & Liu, Z. (2021). Human pose estimation and its application to action recognition: A survey. Journal of Visual Communication and Image Representation, 76, 103055.
[5] Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. Neurocomputing, 312, 135-153.
[6] Lei, M., & Wang, X. (2024). EPPS: Advanced Polyp Segmentation via Edge Information Injection and Selective Feature Decoupling. arXiv preprint arXiv:2405.11846.
[7] Zhou, L., Chen, Y., Gao, Y., Wang, J., & Lu, H. (2020). Occlusion-aware siamese network for human pose estimation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16 (pp. 396-412). Springer International Publishing.
[8] Zhang, T., Huang, B., & Wang, Y. (2020). Object-occluded human shape and pose estimation from a single color image. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7376-7385).
[9] Jiang, J., Ji, Y., Wang, X., Liu, Y., Wang, J., & Long, M. (2021). Regressive domain adaptation for unsupervised keypoint detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6780-6789).
[10] Lin, Q., Yang, L., & Yao, A. (2023). Cross-domain 3d hand pose estimation with dual modalities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17184-17193).
[11] Peng, X., Tang, Z., Yang, F., Feris, R. S., & Metaxas, D. (2018). Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2226-2234).
[12] Chen, X., Wang, S., Wang, J., & Long, M. (2021, July). Representation Subspace Distance for Domain Adaptation Regression. In ICML (pp. 1749-1759).
[13] Pan, S. J., Tsang, I. W., Kwok, J. T., & Yang, Q. (2010). Domain adaptation via transfer component analysis. IEEE transactions on neural networks, 22(2), 199-210.
[14] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[15] Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2013). 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.
[16] Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I., & Schmid, C. (2017). Learning from synthetic humans. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 109-117).
[17] Johnson, S., & Everingham, M. (2010, August). Clustered pose and nonlinear appearance models for human pose estimation. In bmvc (Vol. 2, No. 4, p. 5).
[18] Long, M., Cao, Y., Wang, J., & Jordan, M. (2015, June). Learning transferable features with deep adaptation networks. In International conference on machine learning (pp. 97-105). PMLR.
[19] Zhang, Y., Liu, T., Long, M., & Jordan, M. (2019, May). Bridging theory and algorithm for domain adaptation. In International conference on machine learning (pp. 7404-7413). PMLR.
[20] Mu, J., Qiu, W., Hager, G. D., & Yuille, A. L. (2020). Learning from synthetic animals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12386-12395).
[21] Kim, D., Wang, K., Saenko, K., Betke, M., & Sclaroff, S. (2022, October). A unified framework for domain adaptive pose estimation. In European Conference on Computer Vision (pp. 603-620). Cham: Springer Nature Switzerland.

[22] Pardoe, D., & Stone, P. (2010, June). Boosting for regression transfer. In Proceedings of the 27th International Conference on International Conference on Machine Learning (pp. 863-870).
[23] Wang, B., Mendez, J., Cai, M., & Eaton, E. (2019). Transfer learning via minimizing the performance gap between domains. Advances in neural information processing systems, 32.
[24] De Mathelin, A., Richard, G., Deheeger, F., Mougeot, M., & Vayatis, N. (2021, November). Adversarial weighting for domain adaptation in regression. In 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 49-56). IEEE.
[25] Anderson, E., Bai, Z., Bischof, C., Blackford, L. S., Demmel, J., Dongarra, J., ... & Sorensen, D. (1999). LAPACK users' guide. Society for industrial and applied mathematics.
[26] Knowles, A., & Yin, J. (2013). Eigenvector distribution of Wigner matrices. Probability Theory and Related Fields, 155, 543-582.
[27] Tompson, J. J., Jain, A., LeCun, Y., & Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. Advances in neural information processing systems, 27.
[28] Xu, T., & Takano, W. (2021). Graph stacked hourglass networks for 3d human pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16105-16114).
[29] Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In Proceedings of the European conference on computer vision (ECCV) (pp. 466-481).
[30] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5693-5703).
[31] Cai, Y., Ge, L., Cai, J., & Yuan, J. (2018). Weakly-supervised 3d hand pose estimation from monocular rgb images. In Proceedings of the European conference on computer vision (ECCV) (pp. 666-682).
[32] Zhou, X., Huang, Q., Sun, X., Xue, X., & Wei, Y. (2017). Towards 3d human pose estimation in the wild: a weakly-supervised approach. In Proceedings of the IEEE international conference on computer vision (pp. 398-407).
[33] Zhou, X., Karpur, A., Gan, C., Luo, L., & Huang, Q. (2018). Unsupervised domain adaptation for 3d keypoint estimation via view consistency. In Proceedings of the European conference on computer vision (ECCV) (pp. 137-153).
[34] Yang, Y., Ren, Z., Li, H., Zhou, C., Wang, X., & Hua, G. (2021). Learning dynamics via graph neural networks for human pose estimation and tracking. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8074-8084).
[35] Xiu, Y., Li, J., Wang, H., Fang, Y., & Lu, C. (2018). Pose Flow: Efficient online pose tracking. arXiv preprint arXiv:1802.00977.
[36] Doering, A., Iqbal, U., & Gall, J. (2018). Joint flow: Temporal flow fields for multi person tracking. arXiv preprint arXiv:1805.04596.
[37] Zhang, J., Zhu, Z., Zou, W., Li, P., Li, Y., Su, H., & Huang, G. (2019). Fastpose: Towards real-time pose estimation and tracking via scale-normalized multi-task networks. arXiv preprint arXiv:1908.05593.
[38] Su, P., Shen, X., & Chen, H. (2022, September). Multiscale Spatial and Temporal Learning for Human Motion Prediction. In International Conference on Artificial Neural Networks (pp. 593-604). Cham: Springer Nature Switzerland.