EasyChair Preprint

№ 14841

A Two-Stage Augmentation Approach for Domain Generalization in 3D Human Pose Estimation

Kayode Sheriffdeen

September 13, 2024

A Two-Stage Augmentation Approach for Domain Generalization in 3D Human Pose Estimation

Author: Kayode Sheriffdeen

Date: September, 2024

Abstract

This article presents a novel two-stage augmentation approach designed to enhance domain generalization in 3D human pose estimation models. 3D human pose estimation, a critical task in computer vision, faces significant challenges when models are applied to unseen domains with different environmental conditions, lighting, and human appearances. Our proposed framework combines traditional data augmentation with feature augmentation techniques to create domain-invariant representations. In the first stage, geometric and color-based transformations are applied to the input data to improve robustness against domain shifts, while the second stage focuses on feature augmentation using adversarial learning and neural network-based transformations to ensure domain-invariant feature extraction. This two-stage approach leads to significant improvements in the model's ability to generalize across diverse domains. Extensive experiments on popular 3D human pose datasets demonstrate the effectiveness of the proposed method, showing improved performance compared to traditional models. The study provides valuable insights into the potential of combining data and feature augmentations for domain generalization in computer vision tasks.

Keywords

3D human pose estimation, Domain generalization, Data augmentation, Feature augmentation, Adversarial learning, Convolutional neural networks (CNNs), Human-computer interaction, Domain-invariant features, Cross-domain robustness, Pose estimation

Introduction

3D human pose estimation is a crucial task in computer vision, finding applications in a wide range of fields such as motion capture, gaming, sports analysis, healthcare, and human-computer interaction. The goal of 3D human pose estimation is to accurately predict the spatial coordinates of key body joints from a 2D image or video. However, despite remarkable advancements in deep learning, achieving robust performance across different domains remains a significant challenge. Models trained on a specific dataset often fail to generalize well when applied to new domains, such as different environments, lighting conditions, or human appearances. This phenomenon is known as the domain generalization problem.

Domain generalization refers to the ability of a model to generalize its learned knowledge to unseen data distributions, particularly when there are domain shifts between training and testing sets. In real-world scenarios, 3D pose estimation models are expected to function across diverse environments and subjects, yet domain shifts introduce significant performance degradation.

Traditional data augmentation techniques have been employed to mitigate this issue by artificially expanding the training dataset to increase its diversity. However, these techniques often fail to capture the complex nature of domain variations. To address this challenge, we propose a **Two-Stage Augmentation Approach** for domain generalization in 3D human pose estimation, which combines traditional data augmentation methods with feature augmentation strategies to create domain-invariant representations. This framework not only improves the model's robustness to domain shifts but also enhances its overall performance in predicting 3D human poses across various domains.

Related Work

3D Human Pose Estimation Techniques

3D human pose estimation has been extensively researched in recent years, with a primary focus on developing deep learning models to predict 3D keypoints from 2D inputs. The most common approach involves using convolutional neural networks (CNNs) to extract features from 2D images, followed by regression or lifting mechanisms to infer 3D coordinates. Recent advancements include the use of heatmap-based regression, skeletal models, and temporal consistency in video-based pose estimation.

However, a key limitation of many of these models is their dependence on the dataset used for training. For instance, models trained on the popular Human3.6M dataset often struggle to generalize well to other datasets like MPI-INF-3DHP or real-world data. This issue arises due to the domain gap between training and testing datasets, which includes differences in camera angles, lighting conditions, human appearances, and motion patterns.

Domain Generalization in Machine Learning

Domain generalization in machine learning refers to the challenge of designing models that can generalize across different domains without requiring domain-specific fine-tuning. Unlike domain adaptation, where models are adapted to target domains through additional training, domain generalization aims to achieve cross-domain robustness using only the source domain data.

In the context of 3D pose estimation, achieving domain generalization is crucial, as models are expected to perform in real-world scenarios where data is often noisy and subject to various domain shifts. Several strategies have been proposed to address domain generalization, such as data augmentation, meta-learning, and domain-invariant feature learning.

Data Augmentation in Deep Learning

Data augmentation is a widely used technique in deep learning to increase the diversity of the training dataset by applying transformations such as rotation, scaling, cropping, and flipping to the original data. In the context of 3D human pose estimation, augmenting the dataset with varied poses, lighting conditions, and viewpoints can help improve the model's ability to generalize to new domains.

However, traditional data augmentation techniques often fail to fully address the domain shift problem, as they do not capture the underlying domain-specific variations in the data. This motivates the need for a more sophisticated approach to augmentation, one that can enhance the model's ability to generalize across domains while preserving the fundamental structure of the data.

Two-Stage Augmentation Approach

Our proposed **Two-Stage Augmentation Approach** aims to address the limitations of traditional augmentation methods by incorporating both data augmentation and feature augmentation strategies. This two-pronged approach is designed to create domain-invariant features while preserving the essential characteristics of the data required for accurate pose estimation.

Stage 1: Data Augmentation for Domain Invariance

The first stage of our approach involves applying traditional data augmentation techniques to increase the diversity of the training dataset and make the model more robust to domain shifts. In this stage, various transformations are applied to the input data, including:

- **Geometric Transformations**: These transformations include scaling, rotation, flipping, and translation, which simulate different camera angles and viewpoints. By augmenting the data with different geometric configurations, the model becomes more robust to changes in camera perspective, a common source of domain shift in 3D pose estimation.
- **Color and Lighting Augmentations**: Varying the color balance, contrast, brightness, and saturation of the input images simulates changes in lighting conditions. This helps the model become less sensitive to variations in illumination across different domains.
- **Occlusion Handling**: In real-world scenarios, human poses are often partially occluded by objects or other people. Introducing synthetic occlusions during training helps the model learn to infer occluded keypoints, making it more robust to occlusions in unseen domains.

The goal of this stage is to make the model less dependent on domain-specific features by exposing it to a wide range of synthetic variations. However, while this stage enhances the model's robustness to visual changes, it may not fully capture the complex domain-specific variations present in the feature space.

Stage 2: Feature Augmentation for Enhanced Domain Generalization

The second stage of our approach focuses on feature augmentation, which involves augmenting the internal representations learned by the model to create domain-invariant features. Feature augmentation can be implemented using several strategies, including adversarial learning, neural network-based feature transformations, and domain-invariant feature learning.

- **Adversarial Learning**: One way to achieve feature augmentation is through the use of adversarial learning, where the model is trained to produce domain-invariant features by fooling a domain discriminator. The domain discriminator is tasked with distinguishing between features from different domains, while the feature extractor is trained to make the discriminator's task more difficult. This encourages the model to learn domain-agnostic features that are useful for 3D pose estimation across different domains.
- **Neural Network-Based Transformations**: Feature augmentation can also be achieved by applying neural network-based transformations to the learned feature representations. For example, adding noise or applying dropout during training helps prevent the model from overfitting to domain-specific features and encourages the learning of more generalizable features.
- **Domain-Invariant Feature Learning**: Another strategy is to explicitly enforce domain invariance by using regularization techniques that minimize the distance between feature distributions across different domains. This ensures that the model learns features that are invariant to domain-specific variations, improving its ability to generalize to unseen domains.

By combining data augmentation with feature augmentation, our two-stage approach enables the model to generalize across a wide range of domains while maintaining high accuracy in 3D human pose estimation.

Framework Architecture and Methodology

Model Architecture

The architecture used in conjunction with the two-stage augmentation approach is based on a convolutional neural network (CNN) designed for 3D human pose estimation. The model consists of several layers for feature extraction, followed by a regression head that predicts the 3D coordinates of the key body joints.

- **Feature Extraction**: The feature extraction layers are designed to capture both low-level and high-level features from the input data. These features are then passed through the feature augmentation stage to ensure that they are domain-invariant.
- **Pose Regression**: The regression head takes the augmented features and predicts the 3D coordinates of the keypoints. The output is a set of 3D joint locations, which are used to reconstruct the human pose.

Loss Functions and Optimization

Several loss functions are employed to optimize both stages of augmentation:

- **Pose Estimation Loss**: This is the primary loss function used to minimize the difference between the predicted 3D joint coordinates and the ground truth. Commonly used loss functions include Mean Squared Error (MSE) and L1 loss.
- **Adversarial Loss**: In the case of adversarial feature augmentation, an adversarial loss is used to train the feature extractor and domain discriminator. The adversarial loss encourages the feature extractor to produce domain-invariant features by minimizing the discriminator's ability to distinguish between domains.
- **Reconstruction Loss**: If feature transformations are applied during feature augmentation, a reconstruction loss is used to ensure that the transformed features retain the necessary information for accurate pose estimation.

Training and Inference Process

During training, the model undergoes both data augmentation and feature augmentation. The data augmentation stage generates diverse synthetic variations of the input data, while the feature augmentation stage ensures that the learned features are domain-invariant. The combined approach improves the model’s ability to generalize across different domains.

During inference, the model applies the learned domain-invariant features to predict the 3D poses of individuals in previously unseen domains, achieving high accuracy despite the presence of domain shifts.

Experiments and Results

Datasets Used

We evaluated the proposed two-stage augmentation approach on several popular 3D human pose estimation datasets, including:

- **Human3.6M**: A large-scale dataset containing 3D human pose annotations for a variety of subjects performing different activities. It is one of the most widely used benchmarks for 3D pose estimation.
- **MPI-INF-3DHP**: A dataset that includes diverse human poses captured in both indoor and outdoor environments, providing a challenge for domain generalization.

Evaluation Metrics

To evaluate the performance of our approach, we used the following metrics:

- **Mean Per Joint Position Error (MPJPE)**: The average Euclidean distance between the predicted and ground truth joint coordinates.
- **Procrustes Aligned MPJPE**: A version of MPJPE where the predicted poses are aligned to the ground truth using Procrustes analysis to eliminate translation and rotation errors.

Comparison with Baseline Models

We compared our two-stage augmentation approach with several baseline models, including those trained without any augmentation and those using only traditional data augmentation techniques. Our results demonstrate that the two-stage approach significantly improves domain generalization performance, reducing MPJPE by a substantial margin compared to the baselines.

Ablation Studies

We conducted ablation studies to analyze the impact of each stage of augmentation on the overall performance. The results show that both data augmentation and feature augmentation contribute to improved generalization, with the combination of the two stages yielding the best results.

Discussion

Impact of the Two-Stage Augmentation on Domain Generalization

The results of our experiments indicate that the two-stage augmentation approach effectively enhances the generalization ability of 3D human pose estimation models. By introducing synthetic variations in both the input data and the feature space, the model becomes more robust to domain-specific variations and performs well across diverse domains.

Challenges and Limitations

While the two-stage approach improves domain generalization, it is not without limitations. One challenge is the increased computational complexity introduced by the feature augmentation stage. Additionally, the approach may not fully capture extremely complex domain shifts that involve changes in human motion patterns or environmental factors.

Potential Solutions for Limitations

Future research could explore the use of generative models, such as Generative Adversarial Networks (GANs), to generate more realistic synthetic data for augmentation. Additionally, incorporating temporal information into the augmentation process could help address domain shifts related to motion patterns.

Future Directions

Extension to Other Computer Vision Tasks

While this study focuses on 3D human pose estimation, the two-stage augmentation approach could be extended to other computer vision tasks that suffer from domain generalization issues, such as object detection and action recognition.

Real-World Applications

The ability to generalize across domains has significant implications for real-world applications, such as virtual reality, human-computer interaction, and autonomous driving, where models must function in a variety of environments.

Conclusion

In this paper, we introduced a two-stage augmentation approach for domain generalization in 3D human pose estimation. By combining traditional data augmentation techniques with feature augmentation strategies, our approach improves the model's ability to generalize across domains while maintaining high accuracy in predicting 3D poses. Our results demonstrate the effectiveness of this approach in reducing the domain gap and enhancing the robustness of 3D pose estimation models. This research lays the foundation for future studies exploring more sophisticated augmentation techniques and their applications in other computer vision tasks.

References

- Peng, Q., Zheng, C., & Chen, C. (2024). A Dual-Augmentor Framework for Domain Generalization in 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2240-2249).
- Bin, Y., Cao, X., Chen, X., Ge, Y., Tai, Y., Wang, C., ... & Sang, N. (2020). Adversarial semantic data augmentation for human pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16* (pp. 606-622). Springer International Publishing.
- Peng, Q., Zheng, C., & Chen, C. (2023). Source-free domain adaptive human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4826-4836).
- Gong, K., Zhang, J., & Feng, J. (2021). Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8575-8584).
- Chen, X., Lin, K. Y., Liu, W., Qian, C., & Lin, L. (2019). Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10895-10904).
- Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., & Zhang, L. (2020). Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5386-5395).