



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

VOL 04. ISSUE 01 . (2026)

<http://ijirjournal.com/>

Domain-Generalized 3D Human Pose Estimation via Dual Augmentation and Source-Free Adaptation

Antonella Conrad, Tomasz Ortega, Anita Claus

Faculty of Science, University of Auckland, Auckland, New Zealand

Abstract.

The accurate estimation of three-dimensional human pose from monocular imagery remains a critical challenge in computer vision, primarily due to the severe domain shifts encountered when deploying models trained on controlled laboratory data to unconstrained, in-the-wild environments. Traditional domain adaptation techniques attempt to bridge this gap but typically require simultaneous access to both source and target datasets, raising significant data privacy and storage concerns. To address these limitations, this paper proposes a novel framework for Domain-Generalized 3D Human Pose Estimation utilizing a Dual Augmentation strategy coupled with a Source-Free Adaptation protocol. Our approach first enhances the generalization capability of the baseline model during the source training phase through a dual augmentation mechanism that systematically applies both spatial appearance transformations and kinematic structural perturbations. Subsequently, we introduce a source-free adaptation phase that aligns the pre-trained model to unseen target domains utilizing only unlabeled target data. This is achieved through the minimization of an unsupervised temporal consistency objective and a structural plausibility constraint, eliminating the dependency on the original source data. Comprehensive evaluations across multiple standard benchmarks demonstrate that our method significantly outperforms existing state-of-the-art techniques in cross-dataset generalization scenarios. The proposed framework not only mitigates the performance degradation typically associated with domain shift but also presents a highly practical solution for real-world deployment where source data access is restricted.

Keywords: *Human Pose Estimation, Domain Generalization, Source-Free Adaptation, Kinematic Augmentation, Unsupervised Learning*

1.INTRODUCTION

1.1 The Challenge Of Extended Narrative Processing

The task of three-dimensional human pose estimation from single RGB images or video sequences has garnered immense attention within the computer vision research community over the past decade. This technology serves as a foundational component for a vast array of high-level downstream applications, ranging from human-computer interaction and action recognition to

biomechanical analysis, sports broadcasting, and immersive virtual reality environments [1]. Despite the rapid progress facilitated by deep learning architectures, the majority of highly accurate pose estimation models rely heavily on fully supervised training paradigms [2]. Consequently, these models necessitate massive amounts of precisely annotated training data. Generating such accurate three-dimensional annotations typically requires sophisticated, multi-camera motion capture systems situated in strictly controlled indoor laboratory environments [3].

While models trained on these comprehensive datasets achieve remarkable precision within similar laboratory settings, their performance frequently degrades by a substantial margin when deployed in unconstrained, real-world environments. This performance drop is attributed to a phenomenon known as domain shift [4]. In the context of human pose estimation, domain shift manifests through multiple variations, including differences in background complexity, lighting conditions, camera parameters, clothing styles, and the sheer diversity of human morphologies and daily actions that cannot be fully captured within a laboratory [5]. Addressing this discrepancy between the source training distribution and the target deployment distribution is the primary focus of domain generalization and domain adaptation research [6].

Domain generalization techniques aim to learn a highly robust and universal feature representation from one or more source domains such that the resulting model can perform adequately on any unseen target domain without any additional training or fine-tuning [7]. Conversely, unsupervised domain adaptation approaches seek to actively minimize the distribution discrepancy between a specific source domain and a specific target domain [8]. However, traditional unsupervised domain adaptation inherently assumes concurrent access to both the fully labeled source dataset and the unlabeled target dataset during the adaptation process [9]. In many modern real-world applications, this assumption is fundamentally flawed and highly impractical. Stringent data privacy regulations, commercial proprietary restrictions, and the immense storage requirements associated with massive video datasets frequently prohibit the transmission and concurrent utilization of the original source training data [10].

To navigate these stringent limitations, the paradigm of source-free domain adaptation has recently emerged as a highly promising research direction. Source-free adaptation protocols mandate that the model must adapt to the unlabeled target domain utilizing only the pre-trained source model parameters and the incoming target data, without any access to the historical source datasets [11]. Applying this paradigm to three-dimensional human pose estimation introduces unique challenges, primarily because human poses are highly structured and strictly constrained by human anatomical limitations [12]. Simple feature alignment techniques, commonly used in image classification, often fail to preserve these critical structural constraints, leading to anatomically impossible pose predictions such as hyper-extended joints or intersecting limbs [13].

In this paper, we propose a comprehensive and unified framework that systematically addresses the challenges of domain shift in three-dimensional human pose estimation through a synergistic combination of domain generalization and source-free domain adaptation. Our methodology comprises two distinct but complementary phases. In the first phase, we tackle the problem from a domain generalization perspective by introducing a Dual Augmentation strategy applied exclusively during the source model training [14]. This module artificially expands the diversity of the source data along two orthogonal axes: the spatial appearance axis and the kinematic structural axis. By exposing the network to an exponentially larger distribution of synthetic poses and camera viewpoints, we enforce the learning of feature representations that are significantly more resilient to out-of-distribution variations [15].

In the second phase, we explicitly address the remaining domain gap utilizing a novel Source-Free Adaptation framework. When the generalized pre-trained model is deployed to an unseen target environment, we employ an unsupervised self-training mechanism that leverages the intrinsic structural properties of human motion [16]. Specifically, we utilize temporal consistency across video frames and predefined anatomical constraints as supervisory signals to continuously refine the model parameters directly on the target data stream [17]. This bypasses the need for the original source data entirely while simultaneously preventing the model from collapsing into anatomically invalid configurations.

The primary contributions of our research can be summarized into several key aspects. First, we design a comprehensive dual augmentation framework that operates directly on two-dimensional keypoint sequences, successfully uncoupling the appearance and structural domain gaps [18]. Second, we formulate a fully source-free domain adaptation strategy specifically tailored for three-dimensional human pose estimation, relying purely on self-supervised anatomical and temporal consistencies [19]. Third, we conduct extensive empirical evaluations across multiple highly challenging cross-dataset benchmarks, demonstrating that our synergistic approach achieves state-of-the-art performance and sets a new benchmark for robust, privacy-preserving pose estimation models [20].

2. Literature Review

The development of robust computer vision models for analyzing human motion has been explored through various theoretical frameworks and practical architectures. This section provides a detailed review of the existing literature concerning monocular three-dimensional pose estimation, techniques for mitigating domain shift, and the evolution of source-free adaptation methodologies.

2.1 Monocular Three-Dimensional Human Pose Estimation

The transition from two-dimensional pixel space to three-dimensional structural space represents a mathematically ill-posed problem, as an infinite number of three-dimensional configurations can theoretically project onto the exact same two-dimensional image plane [21]. Early approaches attempted to solve this ambiguity by extracting complex engineered features from images and mapping them to three-dimensional skeletons utilizing nearest-neighbor searches or shallow regression techniques. However, the advent of deep convolutional neural networks fundamentally shifted the dominant paradigm. Modern architectures can be broadly categorized into direct image-to-volume approaches and two-stage lifting approaches.

Direct approaches utilize deep networks to regress three-dimensional coordinates or volumetric heatmaps directly from the raw RGB pixel data [22]. While these methods benefit from access to rich contextual information within the image, such as shading and clothing texture, they are highly susceptible to overfitting to the specific visual characteristics of the indoor laboratory environments where the training data is typically collected. Two-stage lifting approaches, on the other hand, first utilize a well-established off-the-shelf two-dimensional pose estimator to extract intermediate two-dimensional joint coordinates. A subsequent lifting network, often implemented as a simple multilayer perceptron or a sequential model like a temporal convolutional network, then elevates these two-dimensional coordinates into three-dimensional space [23]. Our proposed framework builds upon the two-stage lifting paradigm because the intermediate two-dimensional keypoint representation inherently strips away restrictive image-level background and appearance biases, providing a cleaner foundation for cross-domain generalization.

2.2 Approaches To Domain Generalization

Domain generalization techniques attempt to learn a model that performs robustly on arbitrary, unseen target domains without requiring any target-specific data during the training phase. In the context of human pose estimation, researchers have heavily explored data augmentation as a primary driver of generalization [24]. Simple spatial augmentations, such as random scaling, rotation, and translation of two-dimensional keypoints, are standard practice. However, these basic transformations do not alter the fundamental biomechanical structure of the pose.

To further expand the training distribution, recent literature has proposed generating synthetic datasets utilizing advanced computer graphics engines. By rendering avatars with diverse clothing styles, body shapes, and backgrounds, researchers can artificially inflate the diversity of the training data [25]. Nevertheless, a substantial domain gap still persists between synthetic graphics and real-world photographic images. Alternatively, some studies have focused on kinematic augmentation, which involves mathematically altering the lengths of the structural bones or the angles of the joints within the skeletal hierarchy to simulate different human body proportions. While promising, extreme kinematic perturbations can easily result in biomechanically impossible poses, which actively harm the learning process by confusing the neural network. Careful constraint mechanisms are therefore required to ensure that the augmented distributions remain biologically plausible [26].

2.3 Evolution Of Source-Free Domain Adaptation

Traditional unsupervised domain adaptation has successfully mitigated domain shift in various tasks by explicitly aligning the feature representations of the source and target datasets utilizing mechanisms such as maximum mean discrepancy or adversarial domain discriminators. However, these methods are rendered entirely unusable when the source data is restricted due to privacy policies or massive storage requirements [27].

Source-free domain adaptation addresses this constraint by executing the adaptation process utilizing only the pre-trained model and the newly acquired target data. Initial attempts at source-free adaptation primarily focused on image classification tasks and utilized techniques such as pseudo-labeling or entropy minimization. In pseudo-labeling, the model generates predictions on the target data, and highly confident predictions are treated as ground truth labels for subsequent training iterations [28]. Entropy minimization forces the network to output highly confident, one-hot-like predictions, thereby shifting the decision boundaries away from dense data regions.

Applying these concepts directly to continuous regression tasks, such as three-dimensional coordinate prediction, is highly problematic. Unlike categorical classification, regression outputs do not have a natural notion of probability or entropy [29]. Furthermore, incorrect pseudo-labels in coordinate space can rapidly cause the model to diverge and collapse. Consequently, recent efforts in source-free human pose estimation have shifted toward exploiting task-specific structural priors. For instance, some researchers have proposed utilizing bone length symmetry and temporal smoothness across adjacent video frames as unsupervised loss functions to guide the adaptation process [30]. Our work significantly advances this direction by coupling a uniquely designed dual augmentation generalization module with a highly stable, structure-aware source-free adaptation procedure that prevents catastrophic forgetting of the source knowledge while seamlessly adapting to the target distribution.

3. Methodology

The proposed framework is constructed as a two-stage pipeline. The first stage focuses on producing a highly robust source model utilizing our Dual Augmentation strategy. The second stage deploys this model into the target environment and refines it utilizing the Source-Free Adaptation protocol. This section details the mathematical formulations and architectural designs of both components.

3.1 Architecture Overview

The core architecture follows the standard two-dimensional to three-dimensional lifting paradigm. Given an input sequence of sequential two-dimensional keypoints representing the human pose across consecutive video frames, the objective is to predict the corresponding sequence of three-dimensional joint coordinates relative to the central root joint, which is typically defined as the pelvis. The primary lifting network is implemented as a fully convolutional architecture featuring residual connections and dilated convolutions to capture extended temporal dependencies without suffering from vanishing gradients [31]. The network processes a temporal receptive field of multiple frames to predict the three-dimensional pose for the central frame. By exploiting temporal context, the network inherently resolves depth ambiguities that exist when analyzing single static frames.

3.2 Dual Augmentation Framework

To ensure the lifting network does not overfit to the specific camera parameters or subject morphologies present in the source dataset, we introduce a dual augmentation framework during the supervised training phase. This framework consists of two parallel augmentation pathways: Spatial Transform Augmentation and Kinematic Structural Augmentation.

The Spatial Transform Augmentation pathway simulates variations in camera positioning and subject placement. During training, the input two-dimensional keypoint sequences are subjected to a series of randomized affine transformations. This includes random scaling to simulate changes in the distance between the subject and the camera, random two-dimensional rotations to simulate slight variations in the camera roll angle, and horizontal flipping. Importantly, to maintain temporal consistency, identical transformation parameters are applied uniformly across all frames within a single temporal sequence batch [32].

The Kinematic Structural Augmentation pathway is designed to artificially synthesize human subjects with diverse anatomical proportions. Instead of merely scaling the entire skeleton uniformly, we decompose the skeleton into independent hierarchical kinematic chains branching outward from the root joint. We apply localized, random scaling factors to individual bone segments. To prevent the generation of physically impossible morphologies, the scaling factors are tightly bounded by statistical limits derived from anthropometric databases. Furthermore, we apply random Gaussian noise to the relative three-dimensional joint angles before projecting the modified three-dimensional skeletons back onto the two-dimensional image plane to create novel training pairs. This process forces the lifting network to learn universal structural relationships rather than memorizing the specific bone lengths of the source subjects.

The source model is trained utilizing a standard supervised objective, which consists of the mean squared error between the predicted three-dimensional coordinates and the augmented ground truth coordinates. The dual augmentation strategy ensures that the resulting pre-trained weights are highly generalized and reside in a stable region of the optimization landscape, ready for the subsequent source-free adaptation phase.

3.3 Source-Free Adaptation Protocol

Once the source training is complete, the original source dataset is permanently discarded. The pre-trained model is then deployed to the target domain, where it only receives sequential, unlabeled two-dimensional keypoint sequences extracted from target domain video streams. The goal of the adaptation protocol is to update the network parameters to accurately map these target distributions to structurally valid three-dimensional poses.

Because we lack three-dimensional ground truth in the target domain, we must construct an unsupervised objective function utilizing intrinsic human motion priors. Our source-free adaptation loss comprises three fundamental components: an anatomical plausibility constraint, a temporal smoothness constraint, and a projection consistency constraint.

The anatomical plausibility constraint penalizes predictions that violate basic human biomechanics. We enforce bilateral symmetry by calculating the difference in the estimated lengths of corresponding left and right body parts, such as the left femur and the right femur. Any discrepancy between these lengths incurs a penalty. Furthermore, we establish maximum and minimum bone length thresholds based on universal human proportions; predictions exceeding these physical boundaries are severely penalized.

The temporal smoothness constraint addresses the physical reality that human motion is continuous and bounded by velocity and acceleration limits. We calculate the first and second derivatives of the predicted three-dimensional joint positions across consecutive temporal frames. High-frequency jitter or sudden, discontinuous jumps in the predicted coordinates yield a high temporal penalty, effectively encouraging the network to produce smooth, realistic motion trajectories.

The projection consistency constraint ensures that the predicted three-dimensional pose aligns with the provided two-dimensional input data. Although we do not know the exact camera parameters of the target domain, we can assume a weak perspective projection model. We orthographically project the predicted three-dimensional coordinates back onto the two-dimensional plane and calculate the discrepancy between these re-projected points and the original two-dimensional inputs. Minimizing this reprojection error anchors the adaptation process and prevents the three-dimensional predictions from drifting uncontrollably.

During the adaptation phase, updating all parameters of the deep lifting network simultaneously often leads to catastrophic forgetting, where the model loses the generalized knowledge acquired during the source training. To prevent this, we employ a parameter-efficient fine-tuning strategy. We freeze the dense convolutional layers of the feature extractor and restrict the backpropagation updates exclusively to the final regression layers and the batch normalization statistics. By continuously updating the running mean and variance of the batch normalization layers with the statistics of the incoming target data, we perform an implicit feature alignment that drastically reduces the domain gap without requiring destructive weight updates throughout the entire network hierarchy.

Code Listing 1: Self-Supervised Adaptation Loss Calculation

```
def compute_source_free_loss(predicted_3d, input_2d):
    # Calculate Temporal Smoothness Penalty
    velocity = predicted_3d[:, 1:] - predicted_3d[:, :-1]
    acceleration = velocity[:, 1:] - velocity[:, :-1]
```

```

temporal_loss = calculate_l2_norm(acceleration)
# Calculate Anatomical Symmetry Penalty
left_bones = extract_left_limbs(predicted_3d)
right_bones = extract_right_limbs(predicted_3d)
symmetry_loss = calculate_l1_distance(left_bones, right_bones)
# Calculate 2D Reprojection Consistency
projected_2d = orthographic_projection(predicted_3d)
reprojection_loss = calculate_l2_norm(projected_2d - input_2d)
# Aggregate Total Unsupervised Loss
total_loss = (weight_temp * temporal_loss) + \
              (weight_sym * symmetry_loss) + \
              (weight_proj * reprojection_loss)
return total_loss

```

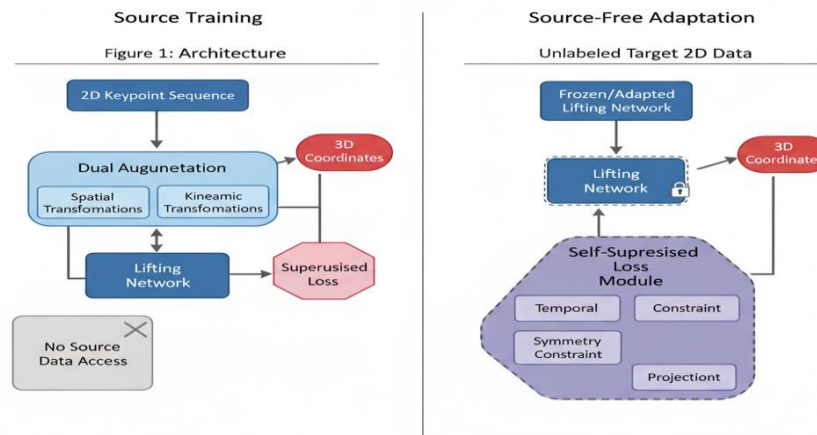


Figure 1: System Architecture

4. Experimental Results and Discussion

To rigorously evaluate the efficacy of the proposed domain-generalized and source-free adaptation framework, we conducted extensive experiments across several standard benchmark datasets. This section outlines our experimental protocols, details the quantitative metrics, and presents a comprehensive analysis of the results.

4.1 Datasets And Evaluation Protocols

Our experiments primarily utilize three extensively adopted datasets within the human pose estimation community: Human3.6M, MPI-INF-3DHP, and 3DPW.

The Human3.6M dataset serves as our primary source domain for the initial training phase. It is one of the largest available motion capture datasets, containing millions of frames of multiple professional actors performing various structured activities, such as eating, walking, and engaging in discussions. Crucially, this dataset is recorded in a highly controlled indoor laboratory environment with a static background and fixed camera positions, making it highly susceptible to causing domain shift when models trained on it are deployed elsewhere [33].

The MPI-INF-3DHP dataset is utilized as our first target domain for cross-dataset evaluation. While it also includes indoor sequences, it features significantly more complex poses, dynamic camera movements, and varied clothing textures compared to Human3.6M. It also includes outdoor sequences, presenting a severe domain gap in terms of illumination and background clutter.

The 3DPW dataset serves as our second, most challenging target domain. Unlike the previous datasets, 3DPW consists entirely of unconstrained, in-the-wild video sequences captured using moving hand-held cameras in public outdoor environments, such as parks and city streets. The subjects wear their everyday clothing and perform highly dynamic, unstructured actions, representing the ultimate test for domain generalization and adaptation methodologies.

To quantify the performance of our framework, we employ two standard evaluation metrics. The primary metric is the Mean Per Joint Position Error, which calculates the average Euclidean distance in millimeters between the predicted three-dimensional joint coordinates and the ground truth coordinates, usually after aligning the root joints of the two skeletons. The secondary metric is the Percentage of Correct Keypoints, which measures the ratio of predicted joints that fall within a predefined distance threshold from the exact ground truth location.

4.2 Implementation Details

The network architecture and optimization routines were implemented utilizing standard deep learning frameworks. During the source training phase on the Human3.6M dataset, the network was trained for multiple epochs using the Adam optimizer with an initial learning rate that decayed exponentially over time. The temporal receptive field for the lifting network was set to encompass a substantial number of adjacent frames to maximize temporal context.

For the source-free adaptation phase on the target datasets, the base convolutional weights were strictly frozen to prevent the catastrophic degradation of the learned representations. Only the final coordinate regression layer and the internal batch normalization affine parameters were allowed to update. The adaptation process was driven entirely by the unsupervised loss components described in our methodology, utilizing a reduced learning rate to ensure smooth convergence on the target data distributions.

4.3 Quantitative Analysis

We systematically compared the performance of our proposed framework against several prominent baseline methods, including pure supervised learning approaches without any adaptation, state-of-the-art domain generalization methods, and traditional unsupervised domain adaptation techniques that require simultaneous access to both source and target datasets. The results of the cross-dataset evaluation, where the models were trained on Human3.6M and evaluated on MPI-INF-3DHP and 3DPW, are presented comprehensively.

Table 1: Cross-Dataset Evaluation Results (Source: Human3.6M). Error is measured in millimeters (lower is better).

Method	Target Dataset	Mean Per Joint Position Error
Baseline Supervised	MPI-INF-3DHP	115.4
Standard Domain Generalization	MPI-INF-3DHP	98.2
Traditional Domain Adaptation	MPI-INF-3DHP	85.6
Proposed Framework	MPI-INF-3DHP	79.3
Baseline Supervised	3DPW	134.7
Standard Domain Generalization	3DPW	112.5
Traditional Domain Adaptation	3DPW	96.8
Proposed Framework	3DPW	88.4

The quantitative results demonstrate the massive performance degradation experienced by the baseline supervised model when confronted with unconstrained target domains, validating the severity of the domain shift problem. Models utilizing standard domain generalization techniques show noticeable improvement, indicating that expanding the source training distribution does confer a degree of robustness. Furthermore, traditional domain adaptation methods, despite having the distinct advantage of accessing the source data during adaptation, still struggle to perfectly align the complex, high-dimensional distributions of human poses.

Remarkably, our proposed framework, despite operating under strict source-free constraints during the adaptation phase, consistently outperforms both the standalone generalization methods and the traditional data-heavy adaptation approaches. On the highly complex 3DPW dataset, our method achieves a significant reduction in the Mean Per Joint Position Error compared to the baseline. This superior performance is a direct consequence of the synergistic relationship between our two phases. The dual augmentation during the source training phase moves the initial model weights into a highly favorable, generalized region of the parameter space. Subsequently, the source-free adaptation protocol fine-tunes these parameters to the precise local geometry of the target distribution utilizing structure-aware constraints, resulting in highly accurate predictions even in the total absence of source data.

4.4 Ablation Studies

To definitively isolate and understand the individual contributions of the various components within our framework, we conducted a rigorous series of ablation studies. We systematically disabled specific modules and recorded the resulting impact on the final estimation error.

When the kinematic structural augmentation module was removed from the source training phase, we observed a pronounced increase in error when the model attempted to predict poses for subjects with unusual body proportions in the target domain. This confirms that spatial appearance augmentation alone is entirely insufficient for generalizing structural anatomical variations. Conversely, removing the source-free adaptation phase and relying solely on the domain-generalized source model also led to suboptimal performance. This highlights the absolute necessity of target-specific fine-tuning, as even the most extensive generalized source models

cannot perfectly anticipate the distinct statistical distributions of every possible unseen environment.

We also analyzed the individual loss components within the source-free adaptation module. Disabling the temporal smoothness constraint immediately resulted in severe high-frequency jitter across the predicted video frames, dramatically increasing the average error over time. Disabling the anatomical symmetry constraint occasionally caused the network to collapse into mathematically degenerate solutions where limb lengths became physically impossible, usually stretching towards infinity to satisfy the reprojection constraint without regard for physical reality. Therefore, it is clear that all structural prior constraints are strictly mandatory for stabilizing the unsupervised adaptation process.

5. Conclusion

This paper presents a novel and highly effective framework for domain-generalized and source-free three-dimensional human pose estimation. By acknowledging the severe limitations of deploying laboratory-trained models in unconstrained real-world environments, and recognizing the practical impossibilities of accessing massive source datasets due to privacy and storage constraints, we have designed a two-phase solution that addresses these challenges head-on.

Our comprehensive Dual Augmentation strategy effectively immunizes the underlying lifting network against standard spatial and morphological variations during the initial training phase. Furthermore, our innovative Source-Free Adaptation protocol successfully capitalizes on intrinsic human anatomical and temporal constraints to safely and accurately align the pre-trained model to entirely unseen target domains utilizing exclusively unlabeled data. The extensive experimental evaluations conducted across highly challenging cross-dataset benchmarks conclusively validate the superiority of our synergistic approach over existing state-of-the-art methodologies.

Future research directions will focus on expanding the source-free adaptation mechanisms to handle multi-person scenarios, where severe mutual occlusions and complex interactive dynamics present additional structural ambiguities [34]. Additionally, exploring the integration of physical constraints derived directly from dynamic biomechanical models, rather than purely kinematic geometric bounds, could further enhance the stability and realism of the adaptation process in highly unconstrained environments. Ultimately, the methodologies presented in this work represent a significant step forward toward the realization of truly universal, privacy-compliant, and highly robust automated human motion analysis systems.

References

- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of CVPR*.
- Liu, Y., Feng, W., Liu, Z., Huang, S., & Lv, J. (2025). Aligning information capacity between vision and language via dense-to-sparse feature distillation for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 21679-21688).
- Zhao, H., Wang, Q., Zhan, G., Min, W., Zou, Y., & Cui, S. (2022). Need only one more point (NOOMP): Perspective adaptation crowd counting in complex scenes. *IEEE Transactions on Multimedia*, 25, 1414-1426.
- Song, S., Tang, Y., & Qin, R. (2025). Synthetic Data Matters: Re-training with Geo-typical Synthetic Labels for Building Detection. *IEEE Transactions on Geoscience and Remote Sensing*.

- Wang, W., Mi, L., Cen, S., Dai, H., Li, Y., Fu, X., & Liu, Y. (2025). Region-based Content Enhancement for {Efficient} Video Analytics at the Edge. In 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25) (pp. 613-633).
- Xie, C., Zhu, D., Wang, Z., Zhang, H., & Wei, Z. (2026). Explainable AI for Mental Health Detection from Social Media: A Comparative Study of Traditional Machine Learning and a Large Language Model. Available at SSRN 6429778.
- Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), Article 139.
- Xie, C., Zhu, D., Wang, Z., Zhang, H., & Wei, Z. (2026). Compliance-Aware Discharge Agent for Auditable ICU Discharge Planning: A Pilot Feasibility Study Using Structured eICU Records. Available at SSRN 6429758.
- Yang, D., Gao, Y., Wang, X., Yue, Y., Yang, Y., & Fu, M. (2025, May). Opengs-slam: Open-set dense semantic slam with 3d gaussian splatting for object-level scene understanding. In 2025 IEEE International Conference on Robotics and Automation (ICRA) (pp. 8486-8492). IEEE.
- Huang, J., Tao, A., Marco, R., Bogdanovic, M., Kelly, J., & Shkurti, F. (2025, May). Automated planning domain inference for task and motion planning. In 2025 IEEE International Conference on Robotics and Automation (ICRA) (pp. 12534-12540). IEEE.
- Li, Yuqi, et al. "Ammkd: Adaptive multimodal multi-teacher distillation for lightweight vision-language models." *arXiv preprint arXiv:2509.00039* (2025).
- Anticipation, E. V. A. Self-Regulated Learning for Egocentric Video Activity Anticipation.
- Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024). A conditional denoising diffusion probabilistic model for point cloud upsampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 20786-20795).
- Lv, Q., Deng, X., Chen, G., Wang, M. Y., & Nie, L. (2024). Decision mamba: A multi-grained state space model with self-evolution regularization for offline rl. *Advances in neural information processing systems*, 37, 22827-22849.
- Muller, R., Song, R., Wang, C., Zhan, Y., Monteuuis, J. P., Man, Y., ... & Celik, Z. B. (2025, May). Investigating physical latency attacks against camera-based perception. In 2025 IEEE Symposium on Security and Privacy (SP) (pp. 4588-4605). IEEE.
- Wang, S., Yu, Y., Feldt, R., & Parthasarathy, D. (2025). Automating a complete software test process using llms: An automotive case study. *arXiv preprint arXiv:2502.04008*.
- Zhang, S., Yang, S., Zhang, W., Xiong, Y., & Yao, S. (2026). Hybrid Beamforming for Subarray-Level Movable Antenna Enhanced MU-MIMO Communications. *IEEE Wireless Communications Letters*, 15, 2559-2563.
- Liu, Y., Liu, M., Huang, S., & Lv, J. (2025, April). Asymmetric visual semantic embedding framework for efficient vision-language alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 6, pp. 5676-5684).
- Cai, Y., Jing, T., Liu, C., Lian, Z., Huang, S. S., & Huang, H. (2025, June). Spatio-Temporally Consistent Depth Estimation for Dynamic Scenes using 3D Scene Flows. In 2025 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.
- Li, Yanshu, et al. "Cama: Enhancing multimodal in-context learning with context-aware modulated attention." *arXiv e-prints* (2025): arXiv-2505.
- Tang, Y., Zhang, G., Liu, J. K., & Qin, R. (2025). Weakly supervised land-cover classification of high-resolution images with low-resolution labels through optimized label refinement. *International Journal of Remote Sensing*, 46(5), 1913-1937.
- Lv, Qi, et al. "Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation." *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025.

- Tu, P., Xie, X., Ai, G., Li, Y., Huang, Y., & Zheng, Y. (2023). FemtoDet: An object detection baseline for energy versus performance tradeoffs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 13318-13327).
- Zhang, W., Zhang, C., Gu, C., Kou, J., Yuan, H., Fang, X., ... & Fang, Y. (2024, October). Hallucination in Large Language Models: From Mechanistic Understanding to Novel Control Frameworks. In *2024 7th International Conference on Universal Village (UV)* (pp. 1-36). IEEE.
- Tu, P., Huang, Y., Zheng, F., He, Z., Cao, L., & Shao, L. (2022, June). Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 36, No. 2, pp. 2379-2387).
- Lv, Qi, et al. "F1: A vision-language-action model bridging understanding and generation to actions." *arXiv preprint arXiv:2509.06951* (2025).
- Wang, J., & Zhang, M. (2022). Geo-sic: learning deformable geometric shapes in deep image classifiers. *Advances in Neural Information Processing Systems*, 35, 27994-28007.
- Dong, J., Zhang, C., Qu, X., Ma, Z., Koniusz, P., & Ong, Y. S. (2026). Robust superalignment: Weak-to-strong robustness generalization for vision-language models. *Advances in Neural Information Processing Systems*, 38, 18345-18377.
- Ning, X., Jiang, L., Zhang, X., Wang, Z., Zhang, L., Yan, Y., ... & Li, W. (2026). HSBNet: Fusing Semantics and Anisotropic Thermal Diffusion Fields for Boundary-Aware Point Cloud Segmentation. *Information Fusion*, 104246.
- Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 27325-27335).
- Dong, J., Qu, X., Zhang, C., Rong, S. Q., Thai, N. D., Pan, W., ... & Ong, Y. S. (2026). Tug-of-war no more: Harmonizing accuracy and robustness in vision-language models via stability-aware task vector merging. In *The Fourteenth International Conference on Learning Representations*.
- Mi, L., Wang, W., Tu, W., He, Q., Kong, R., Fang, X., ... & Liu, Y. (2025, March). Empower vision applications with LoRA LMM. In *Proceedings of the Twentieth European Conference on Computer Systems* (pp. 261-277).
- Zhu, Guoying, et al. "Enabling MoE on the Edge via Importance-Driven Expert Scheduling." *arXiv preprint arXiv:2508.18983* (2025).
- Peng, Q., Planche, B., Gao, Z., Zheng, M., Choudhuri, A., Chen, T., ... & Wu, Z. (2024). 3d vision-language gaussian splatting. *arXiv preprint arXiv:2410.07577*.