Lifelong Domain Adaptive 3D Human Pose Estimation

Qucheng Peng[1], Hongfei Xue[2], Pu Wang[2], Chen Chen[1]
[1]Center for Research in Computer Vision, University of Central Florida, Orlando, Florida, USA
[2]Department of Computer Science, University of North Carolina at Charlotte, Charlotte, North Carolina, USA
qucheng.peng@ucf.edu, hongfei.xue@charlotte.edu, pu.wang@charlotte.edu, chen.chen@crcv.ucf.edu

Abstract

3D Human Pose Estimation (3D HPE) is vital in various applications, from person re-identification and action recognition to virtual reality. However, the reliance on annotated 3D data collected in controlled environments poses challenges for generalization to diverse in-the-wild scenarios. Existing domain adaptation (DA) paradigms like general DA and source-free DA for 3D HPE overlook the issues of non-stationary target pose datasets. To address these challenges, we propose a novel task named lifelong domain adaptive 3D HPE. *To our knowledge, we are the first to introduce the lifelong domain adaptation to the 3D HPE task.* In this lifelong DA setting, the pose estimator is pretrained on the source domain and subsequently adapted to distinct target domains. Moreover, during adaptation to the current target domain, the pose estimator cannot access the source and all the previous target domains. The lifelong DA for 3D HPE involves overcoming challenges in adapting to current domain poses and preserving knowledge from previous domains, particularly combating catastrophic forgetting. We present an innovative Generative Adversarial Network (GAN) framework, which incorporates 3D pose generators, a 2D pose discriminator, and a 3D pose estimator. This framework effectively mitigates domain shifts and aligns original and augmented poses. Moreover, we construct a novel 3D pose generator paradigm, integrating pose-aware, temporal-aware, and domain-aware knowledge to enhance the current domain's adaptation and alleviate catastrophic forgetting on previous domains. Our method demonstrates superior performance through extensive experiments on diverse domain adaptive 3D HPE datasets.

Introduction

3D Human Pose Estimation (3D HPE) involves predicting the 3D coordinates of human joints from images or videos, providing a crucial foundation for applications such as person re-identification (Su et al. 2017), action recognition (Lu et al. 2023; Yan et al. 2023; Peng et al. 2025a), virtual reality (Guzov et al. 2021; Yi et al. 2023; Peng et al. 2025b). The 2D-to-3D lifting paradigm (Pavllo et al. 2019; Zheng et al. 2021; Zhao et al. 2023; Peng, Zheng, and Chen 2024), which predicts 3D poses based on 2D poses (Peng, Zheng, and Chen 2023; Peng et al. 2025c), stands as the most widely adopted pipeline in 3D HPE. Despite its significance, annotated 3D data are typically collected in controlled laboratory settings, featuring indoor environments and a limited range of actions performed by a few individuals. Consequently, pose estimators trained on such labeled datasets encounter difficulties in generalizing to diverse in-the-wild scenarios. Thus, the concept of Domain Adaptation (DA) for 3D HPE (Gholami et al. 2022; Chai et al. 2023; Liu et al. 2023) becomes imperative, aiming to integrate knowledge from labeled (**source**) data into a pose estimator capable of effective generalization on unlabeled (**target**) data.

Existing adaptation settings in 3D HPE fail to account for the evolving nature of pose distributions. While source-free domain adaptation methods (Guan et al. 2022; Nam et al. 2023) enable co-training with all target poses, they assume static distributions and do not address realistic distribution shifts. In practice, target pose distributions are inherently non-stationary—*they continuously evolve due to changing environments and the natural variability of individuals performing actions.* This is particularly evident in autonomous driving scenarios, where pose estimators must adapt across diverse contexts: predicting pedestrian intentions in outdoor environments, monitoring passenger safety inside vehicles.

To address these challenges, we propose a novel task: **lifelong domain adaptive 3D human pose estimation**, as shown in Fig. 1b. *To our knowledge, this is the first time a lifelong DA setting has been introduced for the 3D HPE task.* Unlike the adaptation settings presented in Fig. 1a, the lifelong approach begins with pretraining a 2D-to-3D pose estimator on 2D-3D pose pairs from the source domain. The model is then sequentially adapted to distinct target domains, one at a time, without access to annotations. During each adaptation phase, the pose estimator cannot reference poses from the source or any previously encountered target domains. Our objective is to develop an estimator that performs effectively across both the current and all previously encountered target domains. *The lifelong adaptation differs from source-free/test-time adaptation, as it does not retain access to data from previous target domains.*

The proposed lifelong domain adaptation framework for 3D HPE addresses two main challenges: adapting the 2D-to-3D lifting pose estimator to new domains and preserving knowledge from previous domains. To counter these issues, we introduce a framework featuring 3D pose gen-

(a) Different Settings for Adaptation

	General Domain Adaptation	Fine-tune	Source-free Domain Adaption	Lifelong Domain Adaptation
Access Source Data	✓	✗	✗	✗
Access Previous Target Data	✓	✗	✓	✗
Access Previous Target Labels	✗	✗	✗	✗
Access Current Target Data	✓	✓	✓	✓
Access Current Target Labels	✗	✓	✗	✗

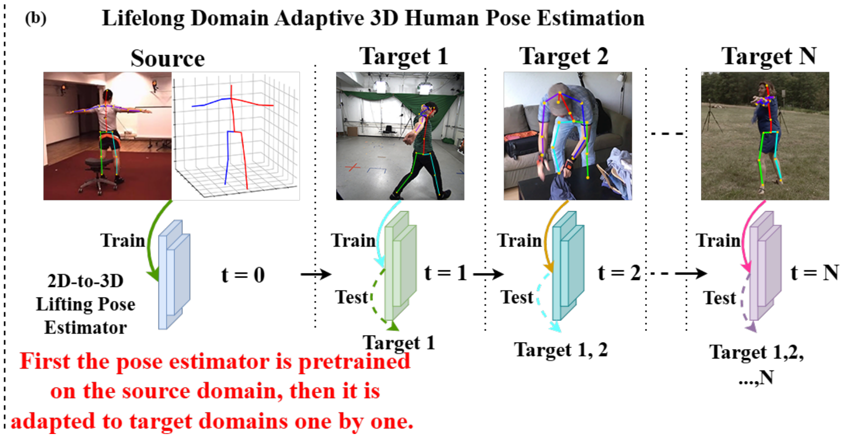

Figure 1: (a) Comparisons among general domain adaptation (Gopalan, Li, and Chellappa 2011), fine-tune (Donahue et al. 2014), source-free/test-time domain adaptation (Sun et al. 2020), and lifelong domain adaptation (Wang et al. 2022a). (b) The paradigm of lifelong domain adaptive 3D human pose estimation.

erators, a 2D pose discriminator, and a 2D-to-3D lifting pose estimator, employing a generative adversarial network (GAN) (Goodfellow et al. 2014) structure to minimize domain shifts. This framework ensures high-quality adaptations by aligning 2D and 3D poses and incorporates a novel 3D pose generator that utilizes pose-aware, temporal-aware, and domain-aware information to enhance adaptation and mitigate catastrophic forgetting. Additionally, a 2D pose diffusion sampler is implemented for efficient domain-aware prior generation. Our contributions can be summarized in three main aspects:

- We introduce lifelong domain adaptive 3D HPE, addressing sequential domain shifts without access to previous domain data. *This is the first work to tackle lifelong domain adaptation in 3D HPE*.
- We demonstrate that mitigating catastrophic forgetting requires synergistic integration of 3D generators, adversarial 2D alignment, and exponential moving average n.
- We conduct comprehensive experiments across multiple benchmarks, demonstrating our approach significantly outperforms existing methods in lifelong adaptation.

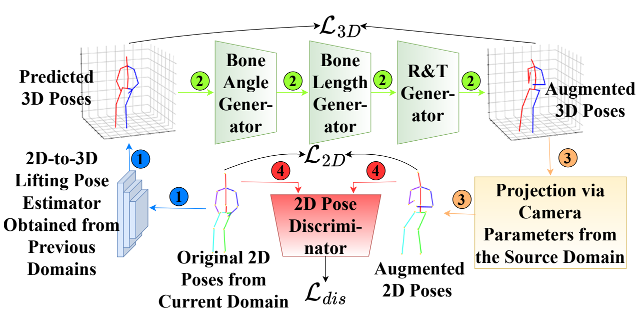

Figure 2: Overall adaption framework of our proposed lifelong domain adaptive 3D HPE approach at time $t = j$.

Related Work

3D Human Pose Estimation. The 2D-to-3D lifting paradigm dominates 3D HPE, where 2D pose estimators (Li et al. 2020b; Wang et al. 2020b) generate predictions that are then transformed into 3D poses. Key approaches include dilated temporal convolutions (Pavllo et al. 2019), transformer-based methods (Zheng et al. 2021), mixed sequence-to-sequence encoders (Zhang et al. 2022), and frequency domain techniques (Zhao et al. 2023).

Lifelong Domain Adaptation. Existing methods focus primarily on classification tasks. TENT (Wang et al. 2020a) minimizes generalization errors through entropy minimization, while CoTTA (Wang et al. 2022b) addresses error accumulation using weight-averaged predictions and stochastic restoration. Other works explore gradual shifts (Marsden, Döbler, and Yang 2022), symmetric measurements (Döbler, Marsden, and Yang 2023), and extensions to object detection (Yang et al. 2022) and person re-identification (Huang et al. 2022). *We present the first lifelong domain adaptation approach for 3D human pose estimation*.

Domain Adaptive 3D Human Pose Estimation. Current DA methods for 3D HPE use both labeled source and unlabeled target data during training. Approaches include GANs for domain discrimination (Gholami et al. 2022), global-local alignment strategies (Chai et al. 2023), multi-hypotheses networks with source augmentation (Liu et al. 2023), and hybrid optimization-learning methods (Jiang et al. 2024). Source-free adaptations combine poses and 3D human shapes (Guan et al. 2022; Nam et al. 2023). *Our work focuses on lifelong domain adaptive 3D HPE, where the estimator processes one pose dataset at a time rather than accessing multiple domains simultaneously*.

Methodology

2D-to-3D Lifting HPE. The predominant paradigm in 3D Human Pose Estimation (3D HPE) involves 2D-to-3D lifting (Pavllo et al. 2019; Zheng et al. 2021; Zhang et al. 2022; Zhao et al. 2019). This paradigm assumes that $x_i^{sr} \in \mathbb{R}^{J\times 2}$

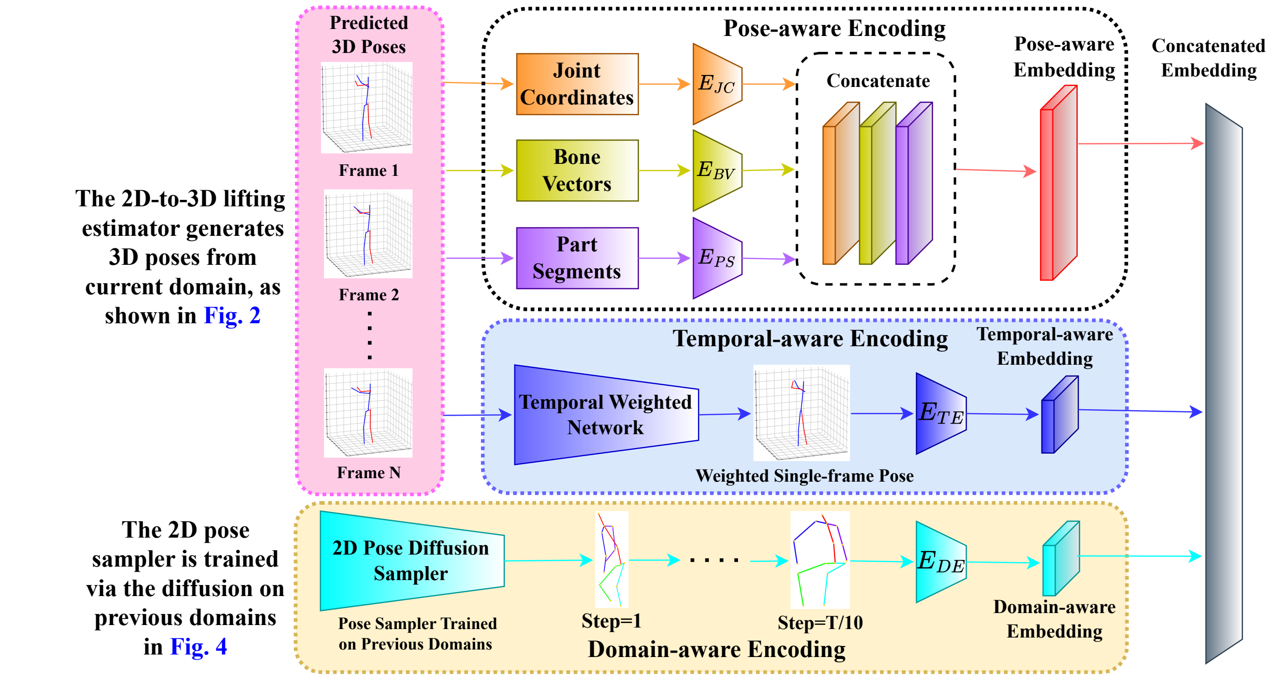

Figure 3: Details of the unified paradigm for each of the three 3D pose generators in Fig. 2. In contrast to existing generators, we introduce part segments (in purple color) for improved pose-aware encoding and introduce additional temporal-aware encoding (in blue color), leading to better adaptation on the current domain. Besides, we employ an unconditional 2D pose diffusion sampler to generate the domain-aware embedding (in cyan color), effectively mitigating catastrophic forgetting.

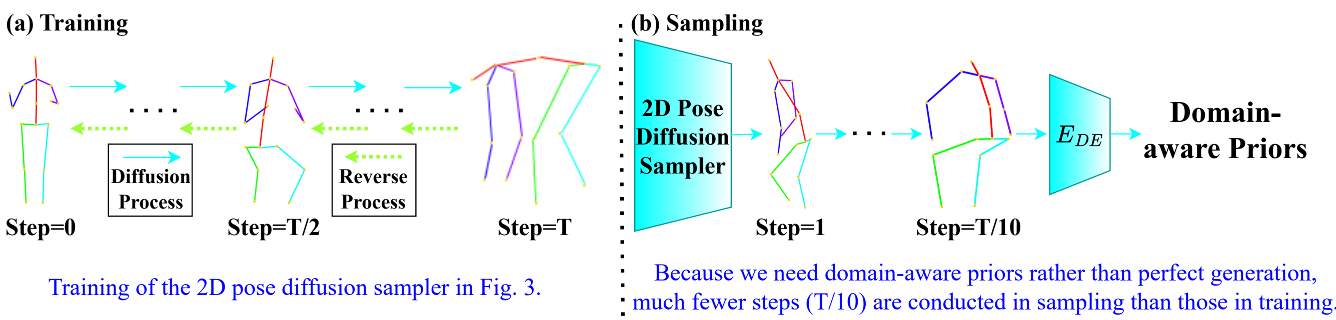

Figure 4: Training (a) and sampling (b) processes of the 2D pose diffusion sampler.

represents the 2D coordinates of J keypoints in a sample from the labeled source domain (with 2D poses as input). Correspondingly, $y_i^{sr} \in \mathbb{R}^{J\times 3}$ denotes the corresponding 3D positions in the camera coordinate system (with 3D poses as output). The source domain, denoted as $sr = \{(x_i^{sr}, y_i^{sr})\}_{i=1}^{M_{sr}}$, contain M_{sr} 2D-3D pose pairs. The 2D-to-3D lifting pose estimator, denoted as $\mathcal{P} : x_i^{sr} \mapsto \hat{y}_i^{sr}$, predicts the corresponding 3D pose positions $\hat{y}_i^{sr}$. However, the fully-supervised paradigm (Pavllo et al. 2019; Zheng et al. 2021) optimized for source poses is inadequate for addressing the DA problem due to the lack of considering domain shifts between source and target domains.

Problem Statement. In our new lifelong setting for 3D HPE, the pose estimator $\mathcal{P}$ is initially pre-trained on the labeled source domain sr. *We define the timestamp for training on source poses as* $t = 0$. Subsequently, it undergoes adaptation to target domains $tg_1, .., tg_N$ sequentially, as shown in Fig. 1(b). For the training at time $t = j \in [1, N]$, $\mathcal{P}$ is exclusively exposed to the target domain $tg_j = \{x_i^{tg_j}\}_{i=1}^{M_j}$. During testing at time $t = j$, the evaluation focuses not only on performance for the current target domain tg_j but also on performance across all previous target domains $tg_1, .., tg_{j-1}$. This approach closely mirrors the handling of non-stationary data in real-world scenarios. However, within the lifelong learning setting, addressing catastrophic forgetting of previous target domains becomes a challenge.

Overview of the Proposed Method. In Fig. 2, we present the comprehensive pipeline of the lifelong 3D HPE for DA at time $t = j$. Given a 2D-to-3D lifting pose estimator $\mathcal{P}$ that has finished the adaptation on previous domains

Time	Method	S5	S6	S7	S8	Avg
t = 0	Source-only	53.5/46.3	56.3/47.7	50.4/41.9	46.6/35.1	-
t = 4	Adaptpose-LL (Gholami et al. 2022)	52.8/46.3	54.3/46.0	47.5/40.9	42.1/30.0	49.2/40.8
	RMT-Pose (Döbler, Marsden, and Yang 2023)	53.0/46.4	54.0/45.5	47.1/40.8	40.9/28.7	48.8/40.4
	CoTTA-Pose (Wang et al. 2022b)	52.2/45.7	53.9/45.1	46.7/40.3	42.0/29.8	48.7/40.2
	CycleAdapt-LL (Nam et al. 2023)	52.5/46.0	54.0/45.5	46.9/40.4	41.7/30.1	48.8/40.5
	PoseDA-LL (Chai et al. 2023)	51.5/44.9	51.9/44.5	46.2/39.5	40.9/28.6	47.6/39.4
	Ours	**48.7/42.5**	**48.6/40.8**	**42.3/36.9**	**40.0/27.4**	**44.9/36.9**

Table 1: Cross-scenario adaptation on H3.6M: S1 $\rightarrow$ S5, S6, S7, S8. MPJPE ($\downarrow$)/PA-MPJPE ($\downarrow$) as metrics.

Time	Method	TS1	TS2	TS3	TS4	TS5	TS6	Avg
t = 0	Source-only	83.0/59.2	105.4/74.5	90.9/60.6	104.7/70.9	109.4/64.5	102.7/77.4	-
t = 6	Adaptpose-LL	69.0/50.1	83.8/55.7	69.4/46.5	82.7/58.2	90.0/59.4	98.5/66.8	82.2/56.1
	RMT-Pose	68.8/49.7	83.5/55.2	69.0/46.3	82.5/57.7	88.6/57.7	97.0/66.2	81.6/55.5
	CoTTA-Pose	68.4/49.5	83.1/54.9	68.5/45.8	81.9/57.2	89.0/58.5	96.6/65.5	81.3/55.2
	CycleAdapt-LL	68.6/49.7	83.5/55.1	68.7/46.2	82.0/57.3	89.4/58.6	96.4/65.4	81.4/55.4
	PoseDA-LL	67.8/48.5	82.4/54.1	67.9/44.8	81.3/56.4	88.5/57.6	96.0/65.3	80.7/54.5
	Ours	**61.1/42.9**	**74.9/49.7**	**62.3/40.9**	**75.4/52.3**	**83.9/55.5**	**94.3/62.6**	**75.3/50.7**

Table 2: Cross-dataset adaptation on H3.6M $\rightarrow$ 3DHP: TS1,...,TS6. MPJPE ($\downarrow$)/PA-MPJPE ($\downarrow$) as metrics.

$tg_1, tg_2, .., tg_{j-1}$, it is initially utilized to predict pseudo 3D poses based on 2D poses from the current domain tg_j. Subsequently, the estimated 3D poses undergo augmentation through three distinctive 3D pose generators, each designed to encode pose-aware, temporal-aware, and domain-aware knowledge, as shown in Fig. 3.

The augmented 3D poses are then projected onto augmented 2D poses using camera parameters from the source domain. The framework's update is facilitated through three distinct loss functions. One is $\mathcal{L}_{3D}$, which aligns predicted 3D poses with augmented 3D poses. Another one is $\mathcal{L}_{2D}$, which compares ground truth 2D poses with augmented 2D poses. Moreover, a 2D pose discriminator is introduced to distinguish between the two types of 2D poses. The min-max game among the 2D pose discriminator and the 3D pose generators is regulated by the third loss $\mathcal{L}_{dis}$.

3D Pose Generation. In Fig. 3, we show the details of our proposed 3D pose generator paradigm illustrated in Fig. 2. Consecutive 3D pose frames estimated by the 2D-to-3D lifting pose estimator serve as the foundation for constructing pose-aware, temporal-aware, and domain-aware embeddings. Projection neural networks $E_{JC}, E_{BV}, .., E_{DE}$ are utilized to project inputs such as joint coordinates or bone vectors to embeddings, and these embeddings are concatenated to form the input for 3D pose generators.

In prior works (Gholami et al. 2022; Chai et al. 2023), pose-aware encoding is achieved by extracting joint coordinates and bone vectors from 3D poses. However, these approaches overlook part-aware information, which is essential for a more comprehensive representation of the human body. Moreover, joints not physically connected or belonging to the same part in the human body model still exhibit relationships that warrant consideration. Consequently, we delineate six body part segments based on the human body: left hand, right hand, left leg, right leg, torso, and extended torso. The first five segments cover the five primary body parts, while the last segment –extended torso– establishes connections for joints that lack physical linkage and do not belong to the same part.

For temporal-aware encoding, multiple consecutive frames of 3D poses are input into a temporal weighted convolutional network, generating a weighted single-frame pose. This process encodes temporal-aware knowledge to enhance the 3D pose generator's synthesis capabilities.

For domain-aware encoding, we introduce domain-aware priors from previous domains as a mitigation strategy against catastrophic forgetting. We choose diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) over GANs (Goodfellow et al. 2014) in terms of providing domain-aware priors due to the ability of diffusion models to preserve mode coverage and diversity, thus avoiding potential mode collapses in GANs (Xiao, Kreis, and Vahdat 2021). This attribute is particularly crucial in the lifelong setting because there is a sequence of distribution shifts among the evolving target poses. Specifically, a 2D pose diffusion sampler, trained on 2D poses from prior domains via Denoising Diffusion Implicit Models (DDIM) (Song, Meng, and Ermon 2020), is employed. The generated 2D poses are subsequently encoded as domain-aware embeddings. Fig. 4 illustrates the training and sampling process.

In the training phase of the 2D pose diffusion sampler, the small size of pose data (typically (16,2) for 16 joints) compared to image size (typically (224, 224) for a standard ResNet (He et al. 2016) input) allows for rapid convergence. During training, the maximum time step is set to T. Following (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020), we pick a random time step k from Uniform[1, T]. During testing, where only domain-aware priors are re-

Time	Method	H3.6M → 3DHP → 3DPW			H3.6M → 3DPW → 3DHP		
		3DHP	3DPW	Avg	3DHP	3DPW	Avg
t = 0	Source-only	96.4/66.5	103.3/63.6	-	96.4/66.5	103.3/63.6	-
t = 2	Adaptpose-LL	90.5/64.1	88.2/50.2	89.4/57.2	80.5/53.4	97.6/62.9	89.1/58.2
	RMT-Pose	90.9/64.3	88.5/50.4	89.7/57.4	79.9/53.3	95.0/61.5	87.5/57.4
	CoTTA-Pose	90.0/63.8	88.7/50.5	89.4/57.2	81.1/53.6	93.2/60.0	87.2/56.8
	CycleAdapt-LL	90.2/64.1	88.3/50.2	89.3/57.2	80.0/53.3	95.9/62.3	88.0/57.8
	PoseDA-LL	88.9/62.1	87.6/49.4	88.3/55.8	79.8/53.0	91.5/53.8	85.7/53.4
	Ours	**75.3/51.1**	**81.7/45.6**	**78.5/48.4**	**78.3/52.2**	**83.7/46.9**	**81.0/49.6**

Table 3: Multi-dataset adaptation on "H3.6M (Source) →3DHP→3DPW" and "H3.6M (Source) →3DPW→3DHP". Values are MPJPE (↓)/PA-MPJPE (↓).

Method	H3.6M → 3DHP → 3DPW		H3.6M → 3DPW → 3DHP	
	3DHP	3DPW	3DHP	3DPW
Ours w/o PS	80.2/56.1	85.2/48.9	79.6/52.9	86.8/49.3
Ours w/o TE	79.6/55.8	83.3/47.4	79.0/52.7	87.1/50.0
Ours w/o DE	83.5/57.4	83.7/47.6	79.2/52.7	88.2/51.7
Ours	**75.3/51.1**	**81.7/45.6**	**78.3/52.2**	**83.7/46.9**

(a) Ablation of 3D pose generators.

Method	H3.6M → 3DHP → 3DPW		H3.6M → 3DPW → 3DHP	
	3DHP	3DPW	3DHP	3DPW
Ours w/o $\mathcal{L}_{2D}$	78.4/54.1	83.9/48.0	80.8/53.5	86.2/48.7
Ours w/o $\mathcal{L}_{3D}$	80.7/56.4	85.4/49.2	82.1/54.0	87.6/49.3
Ours w/o $\mathcal{L}_{dis}$	82.5/58.3	85.8/49.4	83.9/57.6	88.0/51.2
Ours w/o EMA	81.2/57.6	83.7/47.8	80.4/52.9	88.5/51.8
Ours	**75.3/51.1**	**81.7/45.6**	**78.3/52.2**	**83.7/46.9**

(b) Ablation of overall framework.

Table 4: Ablation study of (a) 3D pose generators and (b) the overall framework when $t = 2$ for the two multi-dataset adaptation tasks "H3.6M→3DHP→3DPW" and "H3.6M→3DPW→3DHP".

quired, a complete sampling with the same time step T is unnecessary. Instead, we sample only $T/10$ steps, ensuring efficient processing. Consequently, we replace the randomly generated noise channels utilized in prior works (Gholami et al. 2022; Chai et al. 2023) with more controllable and domain-aware priors for the generators. This substitution proves beneficial in mitigating catastrophic forgetting.

By concatenating these diverse embeddings, we can construct more part-aware, temporal-aware, and domain-aware generators tailored for our lifelong DA tasks in 3D HPE.

Optimization Process. In this paragraph, we discuss the optimization process of our proposed method depicted in Fig. 2. Assuming $t = j \in [1, N]$ (corresponding to the current domain tg_j), for a 2D pose $x_i^{tg_j} \in tg_j$, we derive the predicted 3D pose $\hat{y}_i^{tg_j} = \mathcal{P}_j(x_i^{tg_j})$ using the 2D-to-3D lifting pose estimator obtained before the initiation of $t = j$. The concatenation of the three generators is indicated as $G = G_{BA} \circ G_{BL} \circ G_{RT}$, where G_{BA}, G_{BL}, and G_{RT} correspond to distinctive operations for bone angles, bone lengths, and rotation and translation. Consequently, the augmented 3D pose is expressed as $\tilde{y}_i^{tg_j} = G(\hat{y}_i^{tg_j})$, which is then subjected to projection via camera parameters from the source domain, resulting in the augmented 2D pose $\tilde{x}_i^{tg_j} = Proj(\tilde{y}_i^{tg_j})$. Subsequently, both the original 2D pose and the augmented 2D pose partake in the discrimination process through the 2D discriminator D.

We integrate the Mean Squared Error (MSE) loss along with the feedback loss (Gong, Zhang, and Feng 2021; Li et al. 2020a), which ensures that the augmentation extent is sufficiently substantial for the 3D loss:

$$\mathcal{L}_{3D}(x_i^{tg_j}) = \mathcal{L}_{MSE}(\hat{y}_i^{tg_j}, \tilde{y}_i^{tg_j}) + \|1 - \exp|\hat{y}_i^{tg_j} - \tilde{y}_i^{tg_j}|_{\ell_1}\|. \tag{1}$$

The $\mathcal{L}_{3D}$ term ensures the similarity between predicted and augmented 3D poses within a reasonable range, allowing the augmented poses to differ from the predictions while adhering to human body constraints.

Regarding the 2D loss, we address the significant scale factor. Fully normalizing the scales of the original 2D pose and the augmented 2D pose could result in a loss of domain-aware knowledge. Conversely, entirely disregarding this factor would complicate the alignment between the two types of 2D poses. Hence, we propose the 2D loss:

$$\mathcal{L}_{2D}(x_i^{tg_j}) = \mathcal{L}_{MSE}(x_i^{tg_j}, \tilde{x}_i^{tg_j}) + |\frac{x_i^{tg_j}}{\|x_i^{tg_j}\|} - \frac{\tilde{x}_i^{tg_j}}{\|\tilde{x}_i^{tg_j}\|}|_{\ell_1}, \tag{2}$$

where the first term preserves the scales, and the second term normalizes the scales, striking a balance between the two considerations.

For the discrimination loss controlling the 3D generation process, we employ Wasserstein GANs with gradient penalties (Gulrajani et al. 2017) in our discrimination process:

$$\begin{aligned}\mathcal{L}_{dis}(x_i^{tg_j}) &= \mathbb{E}[D(x_i^{tg_j})] - \mathbb{E}[D(\tilde{x}_i^{tg_j})] \\ &\quad + \alpha\mathbb{E}(1 - \|\nabla_{k_i^{tg_j}} D(k_i^{tg_j})\|),\end{aligned} \tag{3}$$

where $k_i^{tg_j} = \epsilon x_i^{tg_j} + (1-\epsilon)\tilde{x}_i^{tg_j}$, with ϵ randomly drawn from $U[0, 1]$, and α serving as a trade-off parameter.

Based on the three proposed losses, the 3D generator G is updated via:

$$\mathcal{L}_G(x_i^{tg_j}) = \mathcal{L}_{3D}(x_i^{tg_j}) - \beta\mathcal{L}_{dis}(x_i^{tg_j}), \tag{4}$$

while both the 2D discriminator D and the pose estimator $\mathcal{P}$ are optimized via:

$$\mathcal{L}_{DP}(x_i^{tg_j}) = \mathcal{L}_{2D}(x_i^{tg_j}) + \gamma\mathcal{L}_{dis}(x_i^{tg_j}). \tag{5}$$

Here, β and γ are hyperparameters to balance the trade-off between different losses. Following optimization, an exponential moving average strategy (EMA) is applied to obtain the pose estimator $\mathcal{P}_{j+1}$ for the next time $t = j + 1$ as:

$$\mathcal{P}_{j+1} = \eta \mathcal{P}_j + (1 - \eta)\hat{\mathcal{P}}_j, \tag{6}$$

where $\mathcal{P}_j$ is the model initialized before $t = j$ begins, and $\hat{\mathcal{P}}_j$ is the model updated after the adaptation on target domain Tg_j is completed. The smoothing coefficient η is set to 0.99. For the 3D pose generator G and the 2D pose discriminator D, the parameters at timestamp $t = j + 1$ are inherited directly after the optimization at timestamp $t = j$.

Experiments

Datasets and Metrics. Our approach is evaluated on three widely used 3D human pose datasets using a 16-keypoint body model and MPJPE/PA-MPJPE metrics. Human3.6M (H3.6M) features 7 indoor subjects (S1, S5-S8, S9, S11); we use S1 as source and S5-S8 as sequential targets. MPI-INF-3DHP (3DHP) contains indoor/outdoor scenes with six test sets (TS1-TS6) used for cross-dataset adaptation. 3DPW provides challenging in-the-wild scenes with 60 sequences. We evaluate on: (1) H3.6M cross-scenario adaptation S1→S5→S6→S7→S8, (2) cross-dataset adaptation H3.6M→TS1→...→TS6, and (3) multi-dataset tasks H3.6M→3DHP→3DPW and H3.6M→3DPW→3DHP.

Implementation Details. We use fully-connected layers for 3D pose generators and 2D pose estimator (VideoPose3D (Pavllo et al. 2019)), and single convolutional layers for projection and temporal weighted networks. Learning rates are 1e-4 (generators/discriminator) and 5e-5 (pose estimator), with $\alpha = 0.35$ and $\beta = \gamma = 2.5$. We employ Adam optimizer (Kingma and Ba 2014) for generators/discriminator and AdamW (Loshchilov and Hutter 2018) for the pose estimator. Training uses batch size 1024 with 27 frames (Gholami et al. 2022; Chai et al. 2023), 40 epochs for source pretraining, and 30 epochs per target domain adaptation.

For the 2D diffusion pose sampler, we use U-Net (Ronneberger, Fischer, and Brox 2015) with batch size 64, Adam optimizer (lr=2e-4), and 10 training epochs with sampling steps in Uniform[1, 400]. *We use 40 sampling steps for pseudo 2D pose generation, prioritizing domain-aware priors over perfect reconstruction for efficiency.*

Baselines. We establish lifelong DA baselines for 3D HPE by adapting existing methods from two categories. First, we extend 3D domain-adaptive HPE methods AdaptPose (Gholami et al. 2022), PoseDA (Chai et al. 2023), and CycleAdapt (Nam et al. 2023) to lifelong settings as AdaptPose-LL, PoseDA-LL, and CycleAdapt-LL, transferring discriminations and augmentations between current and previous domain poses. Second, we adapt lifelong DA methods RMT (Döbler, Marsden, and Yang 2023) and CoTTA (Wang et al. 2022b) to 3D HPE as RMT-Pose and CoTTA-Pose, replacing classification losses with MSE losses.

Quantitative Results. Tables 1-3 demonstrate our method's superior performance across all benchmarks. Our approach consistently outperforms PoseDA-LL (Chai et al. 2023), achieving average improvements of 2.7mm/2.5mm (MPJPE/PA-MPJPE) at $t = 4$ (Table 1) and 5.4mm/3.4mm at $t = 6$ (Table 2). Notably, we achieve substantial gains of 6.7mm/5.6mm on TS1 at $t = 6$. In the challenging multi-dataset adaptation scenario (Table 3), our method surpasses PoseDA-LL by 9.8mm/7.4mm on average, with particularly strong performance on 3DHP (13.6mm/11.0mm improvement at $t = 2$). These quantitative results underscore our framework's effectiveness in both current domain adaptation and catastrophic forgetting mitigation across diverse lifelong learning scenarios.

Ablation Study on 3D Pose Generators. In Tab. 4a, we perform an ablation study on the components of 3D pose generators. Eliminating the joint coordinates and bone vectors employed in previous works (Gholami et al. 2022; Chai et al. 2023), our focus is on three key components: part segments (**PS**), temporal-aware embedding (**TE**), and domain-aware embedding (**DE**). Based on the results, we observe that **DE** emerges as the most crucial component for preserving knowledge from previous domains. The removal of **DE** results in a degradation of 8.2mm and 6.3mm on the MPJPE of 3DHP and 3DPW, respectively, based on the task "H3.6M→3DHP→3DPW".

Ablation Study on the Overall Framework. Tab. 4b evaluates each component: $\mathcal{L}_{2D}$, $\mathcal{L}_{3D}$, $\mathcal{L}_{dis}$, and EMA. EMA prevents catastrophic forgetting (5.9mm/6.5mm drops when removed), $\mathcal{L}_{2D}$/$\mathcal{L}_{3D}$ enable current domain adaptation, and $\mathcal{L}_{dis}$ preserves knowledge while adapting (7.2mm/4.1mm increases when removed). All components are essential.

Analysis of 2D Pose Sampler's Sampling Steps. Due to the lifelong setting, it is necessary to incorporate knowledge from previously learned domains, and that is why we propose the 2D pose sampler to generate domain-aware priors. In such a case, it is meaningful to investigate how many steps of sampling are optimal for providing these priors. Based on the maximum training steps T= 400, we evaluate several values for sampling steps in Tab. 5.

Analysis of 3D Pose Generation Method. In this paper, we utilize GAN (Goodfellow et al. 2014) for the interpretable generation of 3D poses. In Table 7 and Table 6, we compare our stage-by-stage generation approach with one-stage generative methods such as VAE (Kingma, Welling et al. 2019) and DDIM (Song, Meng, and Ermon 2020). The results highlight the superiority of GAN over other generative models in the context of lifelong domain adaptation for 3D HPE.

Qualitative Results. Qualitative results are depicted in Fig. 5 and Fig. 6. We include Source-only, CoTTA-Pose, PoseDA-LL, Ours, and Ground Truth for qualitative comparisons. It is evident from the visual comparisons that our method outperforms other baselines significantly.

Conclusion

In this study, we propose lifelong domain adaptive 3D human pose estimation to address non-stationary target datasets in real-world scenarios. Our framework comprises 3D pose generators, a 2D pose discriminator, and a pose estimator that leverage a GAN structure to mitigate domain shifts through a min-max game. We introduce a novel 3D pose generator paradigm incorporating pose-aware,

	H3.6M → 3DHP → 3DPW		H3.6M → 3DPW → 3DHP	
Sampling Steps	3DHP	3DPW	3DHP	3DPW
step = 0 (Random Noise)	83.4/57.3	83.7/47.6	79.2/52.7	88.0/51.5
steps = T / 40	80.4/55.7	85.2/48.3	82.9/55.7	87.3/50.6
steps = T / 20	76.8/52.6	82.2/46.1	79.8/53.5	84.2/47.4
steps = T / 10 (Ours)	**75.3/51.1**	**81.7/45.6**	**78.3/52.2**	**83.7/46.9**
steps = T / 5	77.0/53.3	82.8/46.5	80.6/54.1	85.0/47.9
steps = T / 2	79.8/54.1	84.5/47.3	82.3/54.9	87.5/49.1
steps = T	80.5/55.8	84.7/48.0	82.0/55.6	88.0/49.9

Table 5: Analysis of the 2D pose sampler steps for the two multi-dataset adaptation tasks H3.6M→3DHP→3DPW" and H3.6M→3DPW→3DHP" when $t = 2$.

Method	S5	S6	S7	S8	Avg
VAE	52.1/46.2	54.2/46.3	48.0/41.3	40.3/28.0	48.7/40.5
DDIM	51.4/45.6	53.6/45.9	46.2/39.8	40.3/28.0	47.9/39.8
GAN (Ours)	**48.7/42.5**	**48.6/40.8**	**42.3/36.9**	**40.0/27.4**	**44.9/36.9**

Table 6: Comparisons of generative models in 3D pose generation on H3.6M: S1 → S5, S6, S7, S8 when $t = 4$

Method	TS1	TS2	TS3	TS4	TS5	TS6	Avg
VAE	69.8/50.4	84.0/55.1	69.7/48.1	81.9/57.4	87.8/56.6	95.3/64.8	81.4/55.4
DDIM	67.1/48.8	82.6/54.5	67.7/47.4	79.2/55.8	86.4/55.9	95.0/63.3	79.7/54.3
GAN (Ours)	**61.1/42.9**	**74.9/49.7**	**62.3/40.9**	**75.4/52.3**	**83.9/55.5**	**94.3/62.6**	**75.3/50.7**

Table 7: Comparisons of generative models in 3D pose generation on H3.6M → 3DHP: TS1, TS2,..., TS6 when $t = 6$

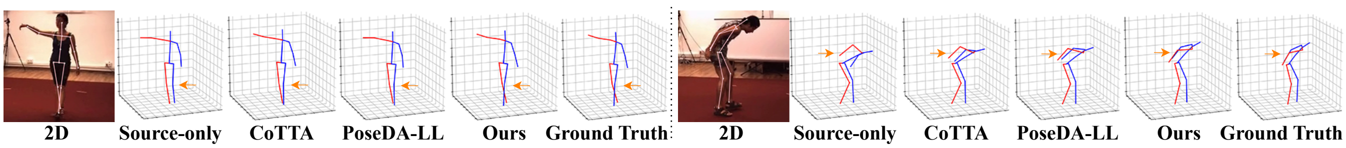

Figure 5: Visualization of H3.6M based on H3.6M: S1 → S5, S6, S7, S8 in Tab. 1. The results are generated via the pose estimator obtained after $t = 4$. Left side is from S5 and right side is from S6.

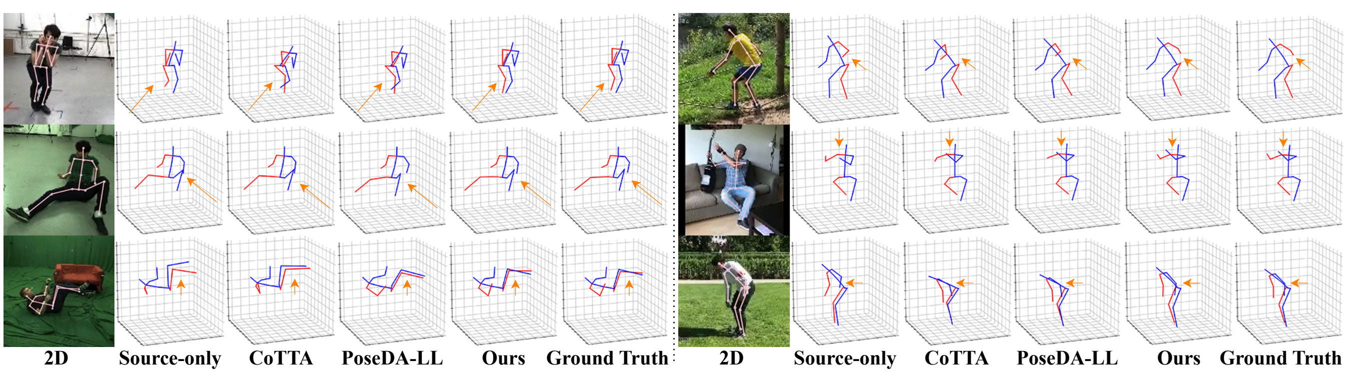

Figure 6: Visualization of 3DHP (left) in Tab. 2 based on H3.6M → 3DHP: TS1,...,TS6 after $t = 6$, and 3DPW (right) based on H3.6M→3DPW→3DHP in Tab. 3 generated after $t = 2$.

temporal-aware, and domain-aware encoding, where pose and temporal components enhance current domain adaptation while the domain-aware component mitigates catastrophic forgetting. Extensive experiments demonstrate significant performance advantages over existing approaches.

References

Chai, W.; Jiang, Z.; Hwang, J.-N.; and Wang, G. 2023. Global Adaptation Meets Local Generalization: Unsupervised Domain Adaptation for 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 14655–14665.

Döbler, M.; Marsden, R. A.; and Yang, B. 2023. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7704–7714.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, 647–655. PMLR.

Gholami, M.; Wandt, B.; Rhodin, H.; Ward, R.; and Wang, Z. J. 2022. Adaptpose: Cross-dataset adaptation for 3d human pose estimation by learnable motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13075–13085.

Gong, K.; Zhang, J.; and 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*, 8575–8584.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gopalan, R.; Li, R.; and Chellappa, R. 2011. Domain adaptation for object recognition: An unsupervised approach. In *2011 International Conference on Computer Vision*, 999–1006.

Guan, S.; Xu, J.; He, M. Z.; Wang, Y.; Ni, B.; and Yang, X. 2022. Out-of-domain human mesh reconstruction via dynamic bilevel online adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 5070–5086.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30.

Guzov, V.; Mir, A.; Sattler, T.; and Pons-Moll, G. 2021. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4318–4329.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, Z.; Zhang, Z.; Lan, C.; Zeng, W.; Chu, P.; You, Q.; Wang, J.; Liu, Z.; and Zha, Z.-j. 2022. Lifelong unsupervised domain adaptive person re-identification with coordinated anti-forgetting and adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14288–14297.

Jiang, Z.; Zhou, Z.; Li, L.; Chai, W.; Yang, C.-Y.; and Hwang, J.-N. 2024. Back to optimization: Diffusion-based zero-shot 3d human pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6142–6152.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; Welling, M.; et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4): 307–392.

Li, R.; Li, X.; Heng, P.-A.; and Fu, C.-W. 2020a. Pointaugment: an auto-augmentation framework for point cloud classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6378–6387.

Li, S.; Ke, L.; Pratama, K.; Tai, Y.-W.; Tang, C.-K.; and Cheng, K.-T. 2020b. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6173–6183.

Liu, H.; He, J.-Y.; Cheng, Z.-Q.; Xiang, W.; Yang, Q.; Chai, W.; Wang, G.; Bao, X.; Luo, B.; Geng, Y.; et al. 2023. PoSynDA: Multi-Hypothesis Pose Synthesis Domain Adaptation for Robust 3D Human Pose Estimation. In *Proceedings of the ACM International Conference on Multimedia*.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Lu, Z.; Wang, H.; Chang, Z.; Yang, G.; and Shum, H. P. H. 2023. Hard No-Box Adversarial Attack on Skeleton-Based Human Action Recognition with Skeleton-Motion-Informed Gradient. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4597–4606.

Marsden, R. A.; Döbler, M.; and Yang, B. 2022. Gradual test-time adaptation by self-training and style transfer. *arXiv preprint arXiv:2208.07736*.

Nam, H.; Jung, D. S.; Oh, Y.; and Lee, K. M. 2023. Cyclic test-time adaptation on monocular video for 3d human mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14829–14839.

Pavllo, D.; Feichtenhofer, C.; Grangier, D.; and Auli, M. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7753–7762.

Peng, Q.; Bai, C.; Zhang, G.; Xu, B.; Liu, X.; Zheng, X.; Chen, C.; and Lu, C. 2025a. NavigScene: Bridging Local Perception and Global Navigation for Beyond-Visual-Range Autonomous Driving. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 4193–4202.

Peng, Q.; Planche, B.; Gao, Z.; Zheng, M.; Choudhuri, A.; Chen, T.; Chen, C.; and Wu, Z. 2025b. 3D Vision-Language Gaussian Splatting. In *The Thirteenth International Conference on Learning Representations*.

Peng, Q.; Zheng, C.; and Chen, C. 2023. Source-free domain adaptive human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4826–4836.

Peng, Q.; Zheng, C.; and 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*, 2240–2249.

Peng, Q.; Zheng, C.; Ding, Z.; Wang, P.; and Chen, C. 2025c. Exploiting Aggregation and Segregation of Representations for Domain Adaptive Human Pose Estimation. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–10. IEEE.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; and Tian, Q. 2017. Pose-driven deep convolutional model for person reidentification. In *Proceedings of the IEEE international conference on computer vision*, 3960–3969.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A.; and Hardt, M. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, 9229–9248. PMLR.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2020a. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020b. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022a. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7201–7211.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022b. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7201–7211.

Xiao, Z.; Kreis, K.; and Vahdat, A. 2021. Tackling the Generative Learning Trilemma with Denoising Diffusion GANs. In *International Conference on Learning Representations*.

Yan, H.; Liu, Y.; Wei, Y.; Li, Z.; Li, G.; and Lin, L. 2023. SkeletonMAE: Graph-based Masked Autoencoder for Skeleton Sequence Pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5606–5618.

Yang, B.; Deng, X.; Shi, H.; Li, C.; Zhang, G.; Xu, H.; Zhao, S.; Lin, L.; and Liang, X. 2022. Continual object detection via prototypical task correlation guided gating mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9255–9264.

Yi, H.; Huang, C.-H. P.; Tripathi, S.; Hering, L.; Thies, J.; and Black, M. J. 2023. MIME: Human-Aware 3D Scene Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12965–12976.

Zhang, J.; Tu, Z.; Yang, J.; Chen, Y.; and Yuan, J. 2022. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13232–13242.

Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; and Metaxas, D. N. 2019. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3425–3435.

Zhao, Q.; Zheng, C.; Liu, M.; Wang, P.; and Chen, C. 2023. PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8877–8886.

Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; and Ding, Z. 2021. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11656–11665.