IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 October 2023, accepted 10 December 2023, date of publication 18 December 2023,
date of current version 28 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3344658

RESEARCH ARTICLE

3DHR-Co: A Collaborative Test-Time Refinement Framework for In-the-Wild 3D Human-Body Reconstruction Task

JONATHAN SAMUEL LUMENTUT[1,2], **(Member, IEEE), AND KYOUNG MU LEE**[1,3], **(Fellow, IEEE)**

[1]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul 08826, South Korea
[2]School of Computer Science, Bina Nusantara University, Jakarta 11530, Indonesia
[3]LG-SNU AI Research, Seoul National University, Seoul 08826, South Korea

Corresponding author: Jonathan Samuel Lumentut (jlumentut@binus.edu)

ABSTRACT The task of 3D human-body reconstruction (3DHR), which mostly utilizes parametric pose and shape representations, has witnessed significant advances in recent years. However, the application of 3DHR techniques in handling real-world in-the-wild data, still faces limitations. Training the 3DHR model in such scenario with 3D human pose's ground truth (GT) is non-trivial. Curating the accurate 3D human pose GT for in-the-wild scenes remains difficult due to various factors. Recent test-time refinement approaches on 3DHR task leverage 2D off-the-shelf human keypoints information to support the lack of 3D supervision on in-the-wild data. However, we observed that additional 2D supervision alone could cause overfitting issue on common 3DHR backbones, making the 3DHR test-time refinement task seem intractable. We answer this challenge by proposing a strategy that complements 3DHR test-time refinement work under a collaborative approach. Specifically, we initially apply a pre-adaptation approach that works by collaborating various 3DHR models in a single framework to directly improve their initial outputs. This approach is then further combined with the test-time adaptation work under specific settings that minimize the overfitting issue to further boost the 3DHR performance. The whole framework is termed as 3DHR-Co, and on the experiment side, the proposed work can significantly enhance the scores of common classic 3DHR backbones up to -34 mm pose error suppression, putting them among the top list on the in-the-wild benchmark data. Such achievement shows that our approach helps unveil the true potential of the common classic 3DHR backbones. Based on these findings, we further investigate various settings on the proposed framework to better elaborate the capability of our collaborative strategy in the 3DHR task.

INDEX TERMS 3D human body reconstruction, test-time refinement, test-time adaptation.

I. INTRODUCTION

3DHR, which has the task of estimating 3D human pose and shape information, has been vastly explored in the computer vision research field. The advanced progress in 3DHR works has been applied to various tasks, including the gaming, health, and fashion industries that employ human pose tracking. Although these 3DHR works [4], [5], [6], [7], [8] have shown remarkable results, their learning process from in-the-wild data is still limited. This issue arises as acquiring the 3D labels from in-the-wild scenario necessitates specialized equipment such as body-mounted sensors [9] or multi-view cameras [10], which incur substantial expenses associated with data collection. Consequently, this scarcity of data poses a significant challenge in achieving robust 3DHR under in-the-wild scenarios. To tackle the issue above, 3DHR works embraced the strategy of non-fully-supervised learning [4], [5], [11]. Prior to this trend, however, early works were mostly focused on the human pose estimation tasks, which the following works [12], [13], [14] handled the few 3D label limitation by feeding the multi-view images input information. In the upcoming years, these works

The associate editor coordinating the review of this manuscript and approving it for publication was Yongjie Li.

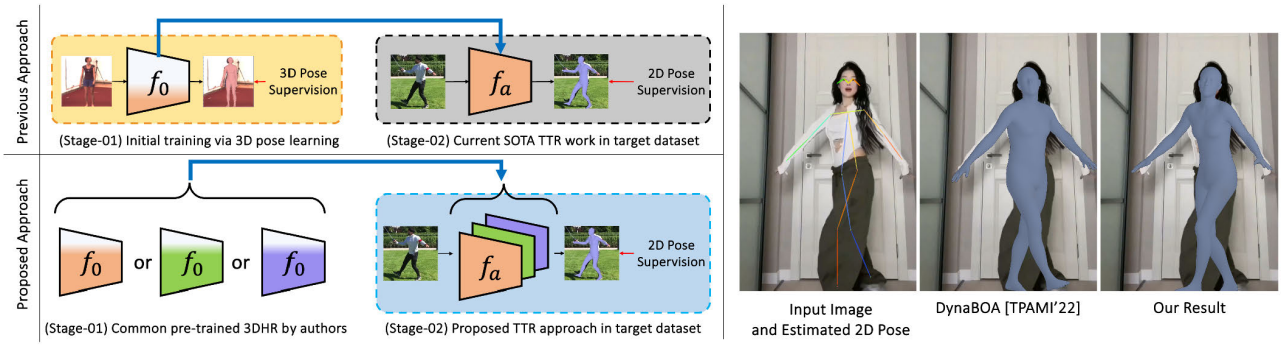

FIGURE 1. ***Left column*: We exhibit the strategy of our approach (lower branch) compared to the recent work (upper branch) in the test-time refinement strategy of 3DHR task. Our approach acts as a refinement tool from given pre-trained 3DHR backbones that the previous works (BOA [1] and DynaBOA [2]) were not intended to (they only focused on the test-time 3DHR domain adaptation task). *Right column*: With given arbitrary input data, our method provided better output than the recent work as it solves the pose ambiguity issue. The TTR term in the figure above stands for test-time refinement while SOTA stands for state-of-the-art.**

by [15], [16], and [17] leveraged the weakly-supervised learning strategy that basically utilizes the non-related 2D and 3D label-based data together for training their human pose estimation network.

As the above's human pose estimation works showed considerable performances, pioneer works [4], [5], [11] are proposed to directly infer the human pose and shape outputs from the image or video input via parametric function (e.g. SMPL [18], MANO [19]). Those parametric-based works are also known for their straightforward benefit during learning as they can utilize the estimated 3D human pose parameters that are projected to 2D representations, to be supervised by the richly-available 2D ground truth dataset (such as the dataset of MSCOCO [20]). These works were then evolved into the temporal strategy (*e.g.*: [8], [21], [22], [23]) that basically utilized features of neighboring frames of video input to subdue the lack of feature information in particular frame or sequences due to *unseen* body parts [24].

As those recent 3DHR works [4], [5], [11] showed considerable ability in predicting human pose and shape outputs by including 2D data supervision, recent approaches, such as BOA [1] and DynaBOA [2] utilized the 2D detected pose information provided from the off-the-shelf detector to firmly guide the 3DHR model during inference-time learning or known as test-time adaptation. Such strategies pave the novel way to achieve top performance in the in-the-wild benchmark data (such as 3DPW [9]) without much modification on the network architecture level. Their works, however, are only focused on the task of domain adaptation. As shown in the left column's upper branch of Figure 1, they (BOA or DynaBOA) require an initial training step with specific dataset (*e.g.*: Human3.6M [25]) that is limited in terms of variety but provides 3D ground truth, to be later adapted on test time with both Human3.6M [25] (as *Source* domain) and in-the-wild data (*e.g.* 3DPW [9] as *Target* domain).

Based on the matter above, our observation in Figure 2 showed that the SPIN's [5] backbone's performance (middle bars of Figure 2) that are directly plugged onto BOA's framework (*BOA-plugged*), was only showing on-par results with the original BOA's work that has its own modified backbone version (left-bars). This is detrimental as *BOA-plugged*, pre-trained with various datasets, should have shown better performance than the original BOA, which was solely trained with Human3.6M [25]. This issue is caused by the overfitting problem on *BOA-plugged*, as it directly adapts the whole video data in the target domain, meaning that any video frame sequences unrelated to the initially fetched frames are insufficient for adaptation.

We tackled the issue above with a straightforward objective: proposing a refinement framework that can plug the original pre-trained 3DHR model f_0 and directly show better refined 3D body results in test time than the initial version via its adapted version f_a. The general idea of this setup is displayed in the left column's lower branch part of Figure 1. The knowledge of each backbone is represented with different colors, and our test-time refinement strategy is meant to increase each's knowledge to predict better 3DHR outputs. Following the idea of BOA (or DynaBOA) that utilizes off-the-shelf 2D pose information during adaptation, our work was able to show better visual results (right column of Figure 1) than the in-the-wild 3DPW [9]'s recent adaptive-based method (DynaBOA [2]).

Our strategy, termed as *3DHR-Co*, is a test-time refinement approach that utilizes currently available 3DHR methods to work collaboratively in distilling the information from one model (as a teacher) unto another (as a learner). This is made possible by treating the learner side of the 3DHR model with the perturbed version while the teacher side with the non-perturbed version of input data during test time. The 3DHR outputs difference between perturbed and non-perturbed versions provide implicit cues for learning during inference time, making it similar to the strategy self-supervised zero-shot learning [26]. In the following sections of this manuscript, we provide further discussion related to our works and the detailed strategies for better understanding. To summarize, our contributions are as follows:

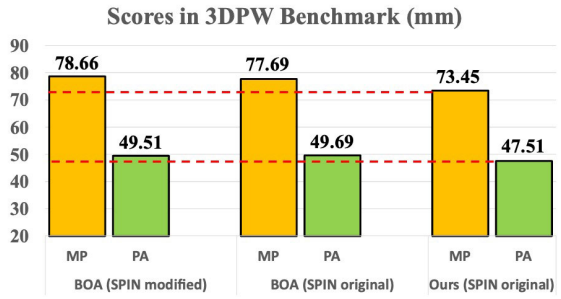

FIGURE 2. The error scores of BOA [1] with original SPIN [5] backbone pre-trained with various datasets (middle) and with modified SPIN backbone pre-trained with single dataset (left) that is on-par. We tackled this issue by performing a refinement strategy that pushed the error score lower (right).

- The proposal of a collaborative-based test-time refinement strategy for 3DHR task (3DHR-Co) that can refine 3DHR models on the go.
- Top-level performance achievement on the in-the-wild benchmark data, obtained by only using the classic common 3DHR backbones run via our collaborative-based test-time refinement framework.
- A study that provides discussions and recommendations regarding the optimal solution for running 3DHR test-time refinement under collaborative strategy.

II. RELATED WORKS

We describe several prior works that are relevant to our main studies. They are classified and described in the following subsections:

A. 3D HUMAN-BODY RECONSTRUCTION

Recently, 3D human-body reconstruction has gained popularity in the computer vision research community. Early works focused on estimating the pose of humans in a stick-man representation, mostly known as the task of human pose estimation [12], [13], [14], [15], [16], [17]. In the past few years, parametric-based approaches have been proposed to simplify the prediction task of 3DHR by estimating the human pose and shape outputs directly. Among these parametric-based approaches, one particular work (namely skinned multi-person linear model or SMPL [18]) was proposed to represent the non-clothed human mesh form with those parameters. Early work termed as Human Mesh Recovery (HMR) [4] utilized deep learning architecture [3] to provide features for regressing the SMPL parameters and camera outputs to render the final human mesh output. The following works then utilized a similar strategy but with several modifications on the backbone levels to support better 3DHR performance from single input image [27], [28]. Realizing that image-based works can be extended to the temporal domain (video-based), various works are also proposed to tackle the limitation of single image input-based 3DHR works [8], [11], [22], [23]. These approaches are naturally better as temporal strategy provides features from neighboring frames that help recognize certain human poses that might be undetected in a particular frame. Just a while ago, recent works [29], [30], [31] involved particular expressive parameters such as hand and face details. Moreover, the current developments [32], [33], [34], [35] also included clothing details information which is pleasing to look on.

B. TEST-TIME LEARNING

Test-time learning, also known as test-time training, is an approach that performs learning on the deep network model during inference time that aims to increase the knowledge of the model itself to be more robust in solving certain task [36]. Other non-3DHR works [26], [37], [38] have experimented with test-time learning to boost their performances. One interesting work, zero-shot image restoration [26], showed that test-time learning is possible from an untrained model. This strategy is further improved by these recent works that mainly utilized the meta-learning approach to firstly train the untrained model and then do the test-time learning procedure for fast adaptation purpose [39], [40], [41]. In recent times, however, the paradigm of test-time learning has intersected with the idea of domain adaptation, which aims to adapt a model learned from a specific domain to be robust on certain target domains. This idea later evolved into domain generalization, where the model is adapted from certain to generalize well to out-of-distribution various target data. In relation to the 3DHR works recent studies of BOA [1] and DynaBOA [2] implemented the idea of test-time training for domain adaptation. DynaBOA [2] utilized additional information retrieval and dynamic adaptation to surpass its predecessor (BOA [1]). These prior works differ from us in terms of motivation, as they focus on the task of solving domain shift issues via test-time adaptation.

III. METHOD

A. PRELIMINARY

In this passage, we describe the general idea of our approach. In contrast to prior works [1], [2], ours is focused on the task of refinement strategy of the pre-trained-available 3DHR models. The objective above led us to the strategy of collaborative learning, where two different 3DHR models are placed directly into our refinement framework. As shown in Figure 3, our framework provided 2 specific branches (white and blue regions) with one model (f_0 placed on the white-colored upper-branch region) acting as the teacher and the other one (f_s placed on the blue-colored lower-branch region) acts as learner. The knowledge of the teacher model with no perturbance case is transferred to the learner model that is perturbed. The perturbed version is provided with noise addition on the image level that acts as a partial occlusion to the human body. By teaching the learner model with the knowledge of the non-perturbed version, it is natural that the learner model gains more improvement during refinement (visualized by the increased green-colored content of f_s in Figure 3). Based on the idea above, we provide a detailed elaboration of our test-time refinement mechanism for the 3DHR task in the following.

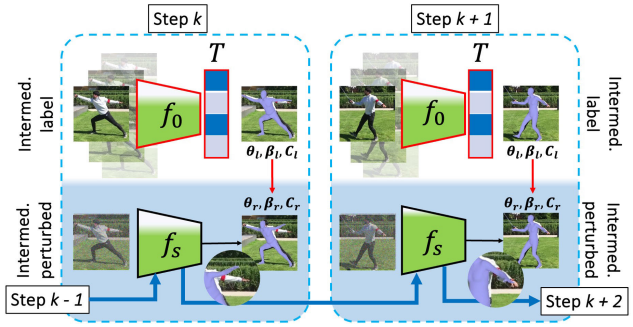

FIGURE 3. Our straightforward yet intuitive test-time refinement framework. The framework above denotes our pre-adaptations strategy (Line 3-11 of Algorithm 1), which aims to provide pre-adapted weight (f_s). The form of f_s is already within our test-time refinement framework; however, it is best to furtherly refine it using bilevel adaptation [1], [2].

For simplicity purposes, we make use of the common 3DHR methods that utilize the SMPL parametric model [18]. The SMPL model is constructed by body pose parameter θ and the shape parameter β. In addition, besides predicting the SMPL parameters, recent common 3DHR methods are also tasked to estimate the camera parameters, c. These SMPL parameters (θ, β) can be used to generate corresponding human mesh representation, $\mathcal{M}$, along with the 3D keypoints representation J via the mesh-to-3D-skeleton mappings provided in SMPL function. The camera parameter c can be used to perform a weak-perspective projection of J from 3D to 2D space. The above nomenclatures are used in the following elaboration.

B. 3DHR-CO TEST-TIME REFINEMENT ALGORITHM

1) MOTIVATION

Our test-time refinement framework is explicitly defined in an algorithm form, and its representation is outlined in the Algorithm 1. As illustrated in Figure 3 earlier, our method is designed to directly refine the pre-trained 3DHR model. To achieve this, we employ two refinement strategies: (i) ***the pre-adaptation stage*** (**Line 3-11** in Algorithm 1), and (ii) ***the full scope*** of our 3DHR-Co test-time refinement, which includes further refinement through our regeneration-based bilevel adaptation (**Line 13-18** in Algorithm 1). The purpose of our pre-adaptation strategy (depicted in Figure 3) is to provide a ready-to-adapt weight that avoids overfitting issues when the bilevel adaptation algorithm is plugged directly with the 3DHR backbone. As bilevel strategy works in a sequential manner, the backbone has the tendency to preserve knowledge on certain sequential frames, leaving the remaining incoming streams to be sub-optimal during adaptation. Our strategy tackled this issue by introducing the sampled target data during the pre-adaptation stage. The effect is crucial, and as demonstrated in Figure 2, our method (right-bars) with BOA function and the pre-adapted SPIN weight can achieve better results than the BOA function with the pre-trained SPIN weight (middle-bars). With this motivation, the whole 3DHR-Co test-time refinement algorithm is constructed to suit the plugged pre-trained model.

Algorithm 1 3DHR-Co Test-Time Refinement Algorithm

Input: Images from benchmark dataset (X)
Require: Pre-trained 3DHR (f_0) and 2D pose guide (G)
Output: Refined outputs ($\hat{\theta}$, $\hat{\beta}$, $\hat{C}$)

1: Create learner model: $f_s = \text{deepcopy}(f_0)$
2: Create batch collections: $B = \text{load}(X)$
3: **while** step $k < K$ and $K \in$ epoch **do**
4: Fetch each data from loader: $B_i = \text{iter\_fetch}(B)$
5: Generate corrupted input: $B_r = \text{corr}(B_i)$
6: Get temporal feats: $V_l = f_0'([B_{i-j}, \ldots, B_i, \ldots, B_{i+j}])$
7: Extract intermediate labels: $(\theta_l, \beta_l, C_l) = T(V_l)$
8: Extract intermediate perturbed: $(\theta_r, \beta_r, C_r) = f_s(B_r)$
9: $\mathcal{L} = \text{loss}(\theta_l, \beta_l, C_l, \theta_r, \beta_r, C_r, G)$ by using Eq. (1)
10: Update learner: $f_s \leftarrow \text{ADAM}(f_s, \nabla_f \mathcal{L})$
11: **end while**
12: Create buffer info: $P = [\,]$
13: **while** batch B is available **do**
14: **if** B.name != P.name **then**
Regenerate new adaptive model: $f_a = f_s$
15: **end if**
16: Extract refined outs: $(\hat{\theta}, \hat{\beta}, \hat{C}, f_a) = \text{Bilevel}(f_a, B, G)$
17: Update buffer P with the latest iteration data: $P = B$
18: **end while**

2) PRE-ADAPTATION STAGE

In this passage, we provide the technical elaboration on running the pre-adaptation strategy. The pre-adaptation strategy is visually depicted in Figure 3 and pseudocode-wise, it is written in the **Line 3-11** in Algorithm 1. To perform pre-adaptation, our work is supported with 2 intermediate output data, namely: test-time ***intermediate label*** and ***intermediate perturbed*** data. The intermediate label data is generated by the non-perturbed version of test input data, while the perturbed version acts as the intermediate perturbed data for the learner module. This approach came with 2 benefits: (a) arbitrary in-the-wild input data can be used directly to extract the perturbed data for refinement purposes, and (b) various models without modification can be learned on the go during test-time. The remaining task is then focused on the strategy of creating a reliable test-time refinement framework.

At the algorithm level, our pre-adaptation approach first fetches the current batch data B_i and its neighboring frames $(B_{i-j}, \ldots, B_{i+j})$ for extracting the temporal features V_l. This procedure is done to extract the intermediate label data (θ_l, β_l, C_l in **Line 7**) extractor, and in this work, the common temporal [8], [23] 3DHR works are directly utilized to fulfill such task. Meanwhile, the intermediate perturbed data (θ_r, β_r, C_r in **Line 8**) version is obtained from the predicted body and camera parameters from test input data B_r corrupted via Gaussian noise (**Line 5**).

This strategy is further learned via MSE loss functionality (**Line 12**):

$$\mathcal{L} = \lambda_1 \|\theta_l - \theta_r\|^2 + \lambda_2 \|\beta_l - \beta_r\|^2 + \lambda_3 \|C_l - C_r\|^2 + \lambda_4 \|G - C_r(J_r)\|^2, \quad (1)$$

and its visual representation is denoted by the red arrow of Figure 3 where each of the output parameters is used for supervision. J_r is the 3D keypoint information extracted from the SMPL output parameters of the perturbed version that are projected into 2D via C_r. The approach above simulates continual knowledge improvement (step-by-step wise) as depicted by the increased green-colored content in the lower-branch scope of the student network (f_s) in Figure 3. Note that the batch of data in each step is sampled randomly. Thus on the step of $k+1$, data can be unrelated to data in the previous step k as shown in Figure 3. Once the pre-adaptation stage is finished, f_a is transferred to the next scope (**Line 13-18**) to perform bilevel test-time refinement.

3) REGENERATION-BASED BILEVEL TEST-TIME REFINEMENT

The full-scope of the 3DHR-Co refinement framework includes the initial pre-adaptation strategy above and the following regeneration-based bilevel refinement approach (**Line 13-18**). The objective of the latter's strategy is to refine the pre-adapted 3DHR backbone's weights while also avoiding the overfitting issue. To run the regeneration strategy, we employ a straightforward weight refreshment mechanism (**Line 14-15**) that runs under frame sequence manner. The data buffer P is utilized to store the latest adapted weight that learns from previous stream sequence data (**Line 17**). For the adaptation, the bilevel function (**Line 16**) is applied, and in this step, f_a is always updated along the batch input. The bilevel function acts as a refinement tool that can adapt the pre-adapted f_a 3DHR model in test-time, and we refer to the original implementation of the respective authors [1], [2].

The regeneration strategy is proposed to avoid the overfitting issue mentioned earlier. The procedure is straightforward (**Line 14-15**) as it re-initiates the model with the weight obtained from the pre-adaptation stage (**Line 3-11**) when a new sequence is fetched. Doing so helps the model to directly re-learn from the pre-adapted version f_s, which already has the knowledge of tested data, rather than the pre-trained version f_0. This step is important as the f_s version already secured initial improvement capability without being overfitted to a particular frame or sequence on test data. These capabilities are further measured in the following *Experiment* section.

IV. EXPERIMENT

A. EXPERIMENTAL SETUP

In this discussion, we provide a comprehensive overview of our implementation method. The pseudocode of our algorithm is implemented using the Pytorch framework. To ensure consistency, various recent and common backbone models are utilized for refinement: SPIN [5] and RSN [27], as they provide the same SMPL parameters and similar Pytorch scripts. The backbones of these works are directly plugged into our algorithm. The intermediate label data extractor is performed via temporal T works of MPN [23] and TCMR [8]. The Eq. 1 is determined with constant coefficients

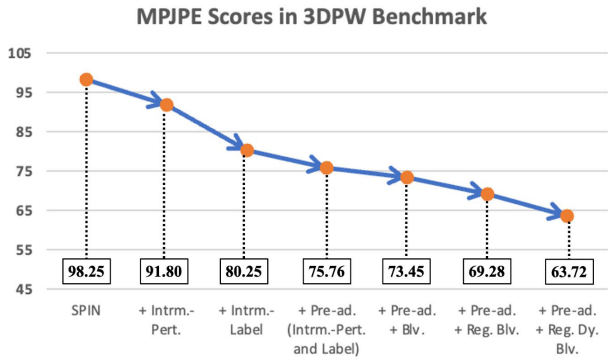

FIGURE 4. Improvement of the classic SPIN [5] backbone in stage-by-stage manner. By accumulating the additional functions proposed in this work, classic SPIN backbone can achieve state-of-the-art result in in-the-wild benchmark data (3DPW [9]). Legends: Intrm.=intermediate, pert.=perturbed, pre-ad.=pre-adaptation, reg.=regeneration, dy.=dynamic, and blv.=bilevel.

($\lambda_1 = 10$, $\lambda_2 = 0.1$, $\lambda_3 = 1$, $\lambda_4 = 1$) while learned via ADAM optimizer with a learning rate of 0.00001 during the pre-adaptation stage (from f_0 to f_s). Our regenerated-based bilevel adaptation function (from f_s to f_a) follows the recent setting of bilevel function of DynaBOA [2]. Our refinement strategy is run on the NVIDIA RTX 3090 GPU, and all experiments are focused on the benchmark of in-the-wild scenario (via 3DPW [9]).

B. FINDING THE WAY FOR IMPROVEMENT

The main challenge of our 3DHR-Co framework is to find reliable data used during test-time learning. As described in the *Method* Section above, we opt to generate the intermediate label from the temporal-based works while the intermediate perturbed data is obtained from the non-temporal models but with noise addition. The benefits above can be seen in Figure 4 where intermediate perturbed data and intermediate label data improve the performance significantly with the MPJPE score of original SPIN **98.25 mm** boosted to **91.80 mm** and **80.25 mm**, respectively, in the 3DPW [9] dataset. Note that MPJPE is a mean per joint position error metric, which means a lower error score translates to better performance. Both findings (intermediate-label and -perturbed) are then applied to our pre-adaptation strategy, and it yields a lower error score (up to **75.76 mm**). With this capability secured, the pre-adapted version is then used in our full-scope strategy along with the bilevel (Reg. Blv.) [1] and dynamic bilevel (Reg. Dy. Blv.) [2] functions. Their scores are shown in Figure 4 accordingly. The recent setting of [2] is utilized in our bilevel function as it shows best score via full-scope 3DHR-Co framework (refer to right-most result with the score of **63.72 mm** in Figure 4). Based on this experiment, we further explore the benefit of both pre-adaptation and the full-scope of 3DHR-Co in the following discussions.

C. DISCUSSIONS ON THE PRE-ADAPTATION TEST-TIME REFINEMENT EXPERIMENT

The effect of our pre-adaptation strategy is showcased in the following experiment. This task is performed by working

FIGURE 5. Qualitative results of SPIN method that are adapted using our pre-adaptation strategy. Significant pose changes are highlighted with red arrow marks. MPN and TCMR act as the intermediate label extractor during pre-adaptation.

TABLE 1. Quantitative scores by performing ablation study on the number of sampled data (percentage) and noise information in our pre-adaptation strategy. In the following, SPIN [5] scores improvement are shown. MPN [23] and TCMR [8] are selected as the intermediate label data generator. Our algorithm allows all methods to work collaboratively in the test-time refinement mechanism. Orange mark defines error gap < -20 mm.

Intrm. -Label \ -Pert.		SPIN	ep-1 (0.07%)			ep-201 (14%)			ep-401 (28%)			ep-601 (42%)		
			σ_1	σ_2	σ_3	σ_1	σ_2	σ_3	σ_1	σ_2	σ_3	σ_1	σ_2	σ_3
Initial		98.25	-	-	-	-	-	-	-	-	-	-	-	-
MPN	MPJPE	-	93.26	94.08	93.26	80.77	82.93	84.36	78.83	79.70	82.44	76.44	77.42	79.92
	Gap	-	-4.99	-4.17	-4.99	-17.48	-15.32	-13.89	-19.42	-18.55	-15.81	-21.81	-20.83	-18.33
TCMR	MPJPE	-	92.90	93.95	94.47	79.86	81.22	83.72	77.72	82.56	82.07	75.76	77.10	79.18
	Gap	-	-5.35	-4.31	-3.79	-18.39	-17.03	-14.53	-20.53	-15.70	-16.18	-22.49	-21.15	-19.07

with the intermediate-label and -perturbed data settings. The intermediate-perturbed data are represented with Gaussian noises varied among $\sigma_1 = 35$, $\sigma_2 = 50$, $\sigma_3 = 65$. The intermediate label version is represented by the number of sampled data (percentage) taken from the tested in-the-wild dataset. In this work, each of the epoch sampled 3 video sequence of the 3DPW benchmark set, where each video contains randomized 8 batch of frames, thus, around 24 frames are utilized from the total 35K frames (0.07%) in 3DPW [9] test set. We empirically determined the maximum number of 600 epochs for further pre-adaptation experiments (42%).

1) WORKING WITH INTERMEDIATE-LABEL AND INTERMEDIATE-PERTURBED DATA

We show the quantitative pre-adaptation scores for both SPIN [5] and RSN [27] models in Table 1 and Table 2 respectively. The scores are shown using MPJPE metric, varied along the number of perturbations in the horizontal direction and the intermediate label extractor (MPN [23] or TCMR [8]) on vertical manner. As shown in the tables above, the more data to be sampled, the larger improvement is achieved as errors are suppressed in a large gap. In the case of SPIN model in Table 1, our pre-adaptation strategy can suppress the error score up to **-22.49 mm** (orange mark at

FIGURE 6. Qualitative results of RSN method that are adapted using our pre-adaptation strategy. Significant pose changes are highlighted with red arrow marks. MPN and TCMR act as the intermediate label extractor during pre-adaptation.

TABLE 2. Quantitative scores by performing ablation study on the number of sampled data (percentage) and noise information in our pre-adaptation strategy. In the following, RSN [27] scores improvement are shown. MPN [23] and TCMR [8] are selected as the intermediate label data meshes generator. Our algorithm allows all methods to work collaboratively in the test-time refinement mechanism. Orange mark defines error gap < −20 mm.

Intrm. Label \ Pert.		RSN	ep-1 (0.07%)			ep-201 (14%)			ep-401 (28%)			ep-601 (42%)		
			σ_1	σ_2	σ_3	σ_1	σ_2	σ_3	σ_1	σ_2	σ_3	σ_1	σ_2	σ_3
Initial		98.51	-	-	-	-	-	-	-	-	-	-	-	-
MPN	MPJPE	-	89.14	86.52	83.58	84.20	85.83	82.90	82.39	83.47	80.37	78.40	85.23	79.93
	Gap	-	-9.37	-11.99	-14.93	-14.32	-12.68	-15.61	-16.12	-15.04	-18.14	-20.11	-13.28	-18.58
TCMR	MPJPE	-	82.70	80.28	83.93	82.13	90.73	84.03	90.14	84.08	85.34	85.09	79.56	83.02
	Gap	-	-15.81	-18.23	-14.58	-16.38	-7.78	-14.48	-8.37	-14.43	-13.17	-13.43	-18.95	-15.49

ep-601), while RSN can be reduced with **-20.11 mm** error gaps (orange mark in Table 2). This indicates that both of these works have hidden potential without modifications at the architecture level. The RSN model case, although more focused on the lower-resolution scale 3DHR task, is still capable of doing test-time refinement, as shown in Table 2. SPIN's qualitative improvement results from its initial version are demonstrated in Figure 5, while RSN's results are shown in Figure 6. The red arrow marks highlighted the significant pose difference between the initial outputs of the pre-trained model and the refined versions via our strategy. MPJPE and PA-MPJPE scores are shown directly below the respective image.

We notice that the appreciable improvements are mainly shown in the body parts that are partially occluded. This phenomenon indicates that our test-time refinement approach in the pre-adaptation stage is already capable of solving the occlusion issue. In a few cases, however, we observed that larger noise (σ_3) might lead to sub-optimal results (shown in the fourth-row image of Figure 5 and Figure 6). Thus, it is recommended to use as low as possible noise value for the 3DPW [9] case. This is in line with our finding as

FIGURE 7. Internet results using our approach that are compared directly with recent top performer of 3DHR adaptation method (DynaBOA [2]). Occluded body parts are plausibly estimated using our approach.

quantitatively, both SPIN and RSN demonstrated the best performances via $\sigma_1 = 35$, as shown in both Table 1 and Table 2, respectively.

D. DISCUSSIONS ON THE FULL SCOPE 3DHR-CO TEST-TIME REFINEMENT EXPERIMENT

The next important step in our experiment is to explore the full-scope version of our 3DHR-Co strategy. Using the pre-adapted weight, our algorithm runs the bilevel function along with the regeneration strategy (**Line 13-18**). This approach is proven to give top performance in the current 3DPW [9] benchmark using only classic SPIN backbone (ResNet-50 [3]). The quantitative scores are shown in Table 3 and Table 4, where MPN and TCMR act as the intermediate label extractor, respectively. Similar to the discussion above, a large error suppression gap (**-34.53 mm**) is obtained by 3DHR-Co(SPIN) with the smallest noise configuration (with σ_1 shown in Table 3). This leads the classic SPIN model to achieve superior performance in the current 3DPW [9] benchmark (MPJPE = **63.72 mm** as highlighted in Table 3) with the MPN as an intermediate label data extractor. Re-plug it with TCMR as an intermediate label data extractor, and the SPIN model is capable of achieving a similar result in the 3DPW benchmark (MPJPE = **63.92 mm** as highlighted in Table 4). The works above conclude that the strategy of using 3DHR intermediate data extractors and the pre-adapted classic backbone together is capable of achieving reliable test-time refinement tasks collaboratively in 3DHR works.

E. DISCUSSIONS ON THE IN-THE-WILD REAL WORLD EXPERIMENT

To further show the validity of our work, 3DHR-Co is demonstrated with real-world internet data cases where no ground truth is available. We follow the DynaBOA setup that first generates the 2D pose via an off-the-shelf 2D human pose estimator (AlphaPose) [42], [43], [44]. Our method is performed via pre-trained SPIN plugged directly into the 3DHR-Co framework with MPN as the intermediate-label extractor. As shown in Figure 7, the stream inputs that have severe occlusions in the human body are recovered better with our approach compared to DynaBOA [2]. The visible errors of [2] (red-arrows in Figure 7) are the issue of human pose

TABLE 3. Quantitative scores of our **full scope 3DHR-Co test-time refinement strategy**. We use DynaBOA to perform the bilevel function (Line 16) in our algorithm. **MPN** is selected as the intermediate label data extractor. **Red** mark below defines error gap < −30 mm.

Intr.-Pert.\-Label	Initial	MPN			
		MP	Gap	PA	Gap
SPIN	98.25 / 60.19	-	-	-	-
3DHR-Co (SPIN)	$\sigma_1 = 35$	**63.72**	-34.53	42.11	-18.08
	$\sigma_2 = 50$	63.87	-34.38	42.10	-18.09
	$\sigma_3 = 65$	63.89	-34.36	42.54	-17.65
RSN	98.51 / 59.81	-	-	-	-
3DHR-Co (RSN)	$\sigma_1 = 35$	64.91	-33.60	42.11	-17.07
	$\sigma_2 = 50$	64.77	-33.74	42.10	-17.08
	$\sigma_3 = 65$	65.19	-33.32	42.54	-16.64

TABLE 4. Quantitative scores of our **full scope 3DHR-Co test-time refinement strategy**. We use DynaBOA to perform the bilevel function (Line 16) in our algorithm. **TCMR** is selected as the intermediate label data extractor. **Red** mark below defines error gap < −30 mm.

Intr.-Pert.\-Label	Initial	TCMR			
		MP	Gap	PA	Gap
SPIN	98.25 / 60.19	-	-	-	-
3DHR-Co (SPIN)	$\sigma_1 = 35$	64.21	-34.04	41.28	-18.91
	$\sigma_2 = 50$	**63.92**	-34.34	41.14	-19.05
	$\sigma_3 = 65$	64.13	-34.12	41.34	-18.85
RSN	98.51 / 59.81	-	-	-	-
3DHR-Co (RSN)	$\sigma_1 = 35$	64.83	-33.68	42.05	-17.13
	$\sigma_2 = 50$	64.82	-33.69	41.97	-17.21
	$\sigma_3 = 65$	65.14	-33.37	41.90	-17.28

ambiguity problem where some occluded paired-body parts are often switched. Our 3DHR-Co approach solved this issue while showing considerable temporal results from the input stream (refer to *supplementary video*).

F. EXECUTION TIME

Our approach came with the extra advantage as it runs with small computational time. Our pre-adaptation approach took approximately 0.15 seconds for a batch of data (each batch contains 8 frames with the size of $3 \times 224 \times 224$ dimensions). For an epoch that sampled 3×8 frames takes around 3 seconds for each pre-adaptation step. As shown in Table 1, running approximately 400 epochs in the pre-adaptation step takes only less than or around half-hour to boost the performance of pre-trained SPIN [5] up to **-21 mm** error suppression on the whole 3DPW [9] test set. Such an approach is favorable compared to re-training a new 3DHR model that takes hours or days to learn. The regeneration-based bilevel steps share similar timing with the original source [2] that take approximately 1 second for adapting a batch of data (1 batch runs 1 frame).

G. LIMITATIONS

While the experiments above show remarkable performances in boosting the 3DHR models, our current development is still limited in terms of boosting test data with the non in-the-wild characteristics. The performance of the full-scope 3DHR-Co framework depends on the intermediate data extraction strategy applied during the pre-adaptation stage. In the pre-adaptation stage, when MPI [45] and Human3.6M [25] are tested directly, our refinement performances are sub-optimal as it only improved with only around **-2 mm** in both cases. We presume this phenomenon happened due to their dataset characteristic that mostly contains homogeneous scenes from in-studio environments. Specifically, their scenes are equipped with static backgrounds and similar color information shared among videos and frames. This condition is well fitted on both networks during the training stage (via [25] training set), and thus, adding noises to such scenes gives minimum effect during test-time refinement. Future studies should consider the constraint above to optimally perform 3DHR refinement in various dataset domains.

V. CONCLUSION

In this work, we propose a collaborative-based test-time refinement framework (termed 3DHR-Co) specialized for boosting a pre-trained 3DHR model during in-the-wild inference scenarios. Performing test-time fine-tuning or refinement on a 3DHR deep neural network model for such scenario remains difficult as the 3D human pose labels are very limited in terms of supervision, while its alternative 2D information may induce an overfitting problem. Our 3DHR-Co framework is proposed to answer this challenge by allowing various common 3DHR models to be boosted directly in test time. This is achieved by employing 2 technical strategies: (i) finding the way of extracting reliable intermediate data during test-time to be used during test-time fine-tuning and (ii) the refinement framework itself that leverages the collaborative strategy. In specific, the collaborative approach involves the process of transferring the knowledge of one 3DHR model to another. Our experimental results showed that even with the common classic 3DHR models, our framework obtained a remarkable performance boost, achieving top performance results and thereby revealing the models' true potential in solving the in-the-wild scenario. The explorations above are also provided with thorough discussions to help find the best test-time refined 3DHR outcomes using our method. We also describe our work limitations for future improvement in 3DHR refinement studies.

REFERENCES

[1] S. Guan, J. Xu, Y. Wang, B. Ni, and X. Yang, "Bilevel online adaptation for out-of-domain human mesh reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10467–10476.

[2] S. Guan, J. Xu, M. Z. He, Y. Wang, B. Ni, and X. Yang, "Out-of-domain human mesh reconstruction via dynamic bilevel online adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 5070–5086, Apr. 2023.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[4] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7122–7131.

[5] N. Kolotouros, G. Pavlakos, M. Black, and K. Daniilidis, "Learning to reconstruct 3D human pose and shape via model-fitting in the loop," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2252–2261.

[6] G. Moon and K. M. Lee, "I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 752–768.

[7] Q. Fang, Q. Shuai, J. Dong, H. Bao, and X. Zhou, "Reconstructing 3D human pose by watching humans in the mirror," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12809–12818.

[8] H. Choi, G. Moon, J. Y. Chang, and K. M. Lee, "Beyond static features for temporally consistent 3D human pose and shape from a video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1964–1973.

[9] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using IMUs and a moving camera," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 601–617.

[10] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt, "Single-shot multi-person 3D pose estimation from monocular RGB," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 120–130.

[11] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3D human dynamics from video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5607–5616.

[12] H. Rhodin, M. Salzmann, and P. Fua, "Unsupervised geometry-aware representation for 3D human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 750–767.

[13] H. Rhodin, F. Meyer, J. Spörri, E. Müller, V. Constantin, P. Fua, I. Katircioglu, and M. Salzmann, "Learning monocular 3D human pose estimation from multi-view images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8437–8446.

[14] Y. Yao, Y. Jafarian, and H. S. Park, "MONET: Multiview semi-supervised keypoint detection via epipolar divergence," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 753–762.

[15] B. Wandt and B. Rosenhahn, "RepNet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7774–7783.

[16] U. Iqbal, P. Molchanov, and J. Kautz, "Weakly-supervised 3D human pose learning via multi-view images in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5242–5251.
[17] B. Wandt, M. Rudolph, P. Zell, H. Rhodin, and B. Rosenhahn, "CanonPose: Self-supervised monocular 3D human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13289–13299.
[18] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.
[19] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–17, 2017, Art. no. 245.
[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[21] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5252–5262.
[22] Z. Luo, S. A. Golestaneh, and K. M. Kitani, "3D human motion estimation via motion compression and refinement," in *Proc. Asian Conf. Comput. Vis.*, 2020.
[23] W.-L. Wei, J.-C. Lin, T.-L. Liu, and H. M. Liao, "Capturing humans in motion: Temporal-attentive 3D human pose and shape estimation from monocular video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13201–13210.
[24] M. Kocabas, C. P. Huang, O. Hilliges, and M. J. Black, "PARE: Part attention regressor for 3D human body estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11107–11117.
[25] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.
[26] A. Shocher, N. Cohen, and M. Irani, "Zero-shot super-resolution using deep internal learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3118–3126.
[27] X. Xu, H. Chen, F. Moreno-Noguer, L. A. Jeni, and F. D. la Torre, "3D human shape and pose from a single low-resolution image with self-supervised learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 284–300.
[28] H. Choi, G. Moon, and K. M. Lee, "Pose2mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 769–787.
[29] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3D deformation model for tracking faces, hands, and bodies," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8320–8329.
[30] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10967–10977.
[31] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, "GHUM & GHUML: Generative 3D human shape and articulated pose models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6183–6192.
[32] S. Saito, Z. Huang, R. Natsume, S. Morishima, H. Li, and A. Kanazawa, "PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2304–2314.
[33] S. Saito, T. Simon, J. Saragih, and H. Joo, "PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 81–90.
[34] T. He, Y. Xu, S. Saito, S. Soatto, and T. Tung, "ARCH++: Animation-ready clothed human reconstruction revisited," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11026–11036.
[35] T. Alldieck, M. Zanfir, and C. Sminchisescu, "Photorealistic monocular 3D reconstruction of humans wearing clothing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1496–1505.
[36] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9229–9248.
[37] T. Ehret, A. Davy, J.-M. Morel, G. Facciolo, and P. Arias, "Model-blind video denoising via frame-to-frame training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11361–11370.
[38] S. Lee, D. Cho, J. Kim, and T. H. Kim, "Restore from restored: Video restoration with pseudo clean video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3536–3545.
[39] S. Park, J. Yoo, D. Cho, J. Kim, and T. H. Kim, "Fast adaptation to super-resolution networks via meta-learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 754–769.
[40] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3513–3522.
[41] S. Lee, M. Choi, and K. M. Lee, "DynaVSR: Dynamic adaptive blind video super-resolution," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2092–2101.
[42] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7157–7173, Jun. 2023.
[43] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "CrowdPose: Efficient crowded scenes pose estimation and a new benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10855–10864.
[44] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2353–2362.
[45] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.

JONATHAN SAMUEL LUMENTUT (Member, IEEE) received the B.Comp.Sc. and M.S. degrees in computer science from Bina Nusantara University, Jakarta, Indonesia, in 2013 and 2014, respectively, and the Ph.D. degree in electrical and computer engineering from Inha University, South Korea, in 2021. From 2022 to 2023, he attended the Postdoctoral Research Program, Seoul National University, Seoul, South Korea. In 2023, he became a Faculty Member of Bina Nusantara University. His research interests include computer vision, human body reconstruction, image restoration, and computational photography.

KYOUNG MU LEE (Fellow, IEEE) received the B.S. and M.S. degrees in control and instrumentation engineering from Seoul National University (SNU), Seoul, South Korea, in 1984 and 1986, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California, in 1993. He is currently with the Department of ECE, SNU, as a Professor. He is an Advisory Board Member of the Computer Vision Foundation (CVF). He has received several awards, particularly the Medal of Merit and the Scientist of Engineers of the Month Award from the Korean Government, in 2018 and 2020, respectively; the Most Influential Paper Over the Decade Award by the IAPR Machine Vision Application, in 2009; the ACCV Honorable Mention Award, in 2007; the Okawa Foundation Research Grant Award, in 2006; the Distinguished Professor Award from the College of Engineering, SNU, in 2009; and the Outstanding Research Award and the Shinyang Engineering Academy Award from the College of Engineering, SNU, in 2010. He has also served as the General Chair for ICCV 2019, ACMMM 2018, and ACCV 2018; the Program Chair for ACCV 2012; the Track Chair for ICPR 2020 and ICPR 2012; and the Area Chair for CVPR, ICCV, and ECCV many times. He has served as an Associate Editor-in-Chief (AEIC) and an Associate Editor for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE; an Associate Editor for the *Machine Vision and Application* journal, the *IPSJ Transactions on Computer Vision and Applications*, and IEEE SIGNAL PROCESSING LETTERS; and an Area Editor for the *Computer Vision and Image Understanding*. He was a Distinguished Lecturer of the Asia–Pacific Signal and Information Processing Association (APSIPA), from 2012 to 2013. More information can be found on his homepage (http://cv.snu.ac.kr/kmlee).