IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 9, 2020, accepted December 14, 2020, date of publication December 18, 2020,
date of current version December 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3045794

Semi-Supervised 3D Human Pose Estimation by Jointly Considering Temporal and Multiview Information

WEI-TA CHU[1], (Senior Member, IEEE), AND ZONG-WEI PAN[2]
[1]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 70101, Taiwan
[2]Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 62102, Taiwan

Corresponding author: Wei-Ta Chu (wtchu@gs.ncku.edu.tw)

This work was supported in part by the Qualcomm Technologies, Inc., under Grant B109-K027D; and in part by the Ministry of Science and Technology, Taiwan, under Grant 108-2221-E-006-227-MY3, Grant 107-2923-E-194-003-MY3, Grant 109-2218-E-002-015, and Grant 107-2627-H-155-001.

ABSTRACT Three-dimensional human pose estimation is usually conducted in a supervised manner. However, because collecting labeled 3D skeletons is expensive and time-consuming, semi-supervised methods that need much fewer amount of labeled 3D data are urgently demanded. Some semi-supervised learning methods propose to independently consider information from consecutive video frames, or frames simultaneously captured from multiple viewpoints. In this article, we propose to jointly consider temporal information and multiview information in a unified adversarial learning framework. Given a 2D skeleton, a pose generator network is developed to estimate the corresponding 3D skeleton, and a camera network is developed to estimate camera parameters. The estimated 3D skeleton is evaluated by a critic network to examine whether the estimated one is a plausible 3D human pose or not. Based on the estimated camera parameters, the estimated 3D skeleton can be re-projected into a 2D skeleton, which should be similar to the input 2D skeleton. The ideas of re-projection and adversarial learning enable the scheme of self supervision. We design network architectures of the aforementioned networks to take 2D skeletons from multiple viewpoints in temporally consecutive frames. By jointly considering two types of information, we verify that performance can be largely improved.

INDEX TERMS 3D human pose estimation, semi-supervised, temporal information, multiview information.

I. INTRODUCTION

Three-dimensional human pose estimation from monocular images has been actively studied in recent years. It is the fundamental step for many advanced research topics, such as human behavior analysis and simulation in virtual reality. Following the success of deep neural networks, and the availability of 3D pose datasets [9], performance of 3D human pose estimation gets impressive recently [6], [13], [15], [17], [20], [22]. Most existing methods attempted to learn a mapping function between monocular images and 3D skeletons in a supervised manner. Given a monocular image, one way to achieve 3D skeleton detection is formulating it as a regression problem. From the image, a heat map representing positions of joints is estimated, and then a 3D skeleton is estimated based on such information. Another way is detecting 2D skeletons first, and then a 2D to 3D skeleton estimation is conducted. No matter which way, labeled 3D skeletons are needed to learn the mapping function. However, manually labeling or correcting 3D skeletons is laborious, making labeled 3D skeletons scarce. In addition, the model learnt in a supervised manner may not be generic to unknown motions and camera positions.

Because of the scarcity of 3D labeled data, some methods have been proposed to find the mapping function in a weakly supervised or a semi-supervised manner. Only a small amount of labeled data are required to guide the initial learning, and self-learning schemes are designed to improve the initial model. For example, in [22], a pre-detected 2D skeleton is input to a pose generator network to generate the corresponding 3D skeleton, and a camera network is developed to estimate camera parameters. The generated 3D skeleton is then re-projected back to a 2D skeleton according to the estimated camera parameters, and the projected one should be similar to the input 2D skeleton.

In our work, we develop semi-supervised 3D human pose estimation by modifying the weakly supervised method proposed in [22]. More importantly, we jointly consider rich

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

information in both temporal and spatial domains. Separately estimating 3D skeletons for each video frame usually yields jiggling. Jointly taking information of consecutive video frames into the estimation model may improve detection robustness. Pavllo *et al.* [17] proposed a temporal convolution model to take 2D skeleton sequences as input to estimate 3D human poses. To utilize rich information from the spatial domain, Chen *et al.* [6] proposed to jointly consider videos captured from two different views. 2D skeletons are first detected from each view. The 2D skeleton from the first view is then synthesized by a network into the second view. The synthesized skeleton should be similar to that originally from the second view. This representation constraint guides network learning, and this network is viewed to be able to describe geometry representation of 3D skeletons.

We integrate the ideas proposed in [17] and [6], and make significant modifications to join temporal information and multiview information into a unified semi-supervised framework, which is trained based on adversarial learning. Although each separate module has been proposed before, how performance can be improved by jointly taking them in a unified model was not investigated before, and this is the main contribution of this article. We will mainly evaluate the proposed framework based on the Human3.6M dataset [9], and verify that significant performance improvement can be obtained.

The rest of this article is organized as follows. Section II presents literature survey on 3D human pose estimation. Section III first briefly describes the semi-supervised adversarial learning framework, and then provides details of how we consider temporal and multiview information. Section IV shows comprehensive experimental studies, followed by conclusion given in Section V.

II. RELATED WORKS

A. CONVENTIONAL 3D HUMAN POSE ESTIMATION

Inferring 3D skeletons from 2D projections can be dated back to the work of Lee *et al.* [11] in 1985. They used lengths of bones and binary decision trees to construct a human pose. Jiang [10] used joint correspondence and searched for optimal 3D human pose from a large amount of data to solve the pose estimation problem. Another way to compile knowledge of 3D human pose is building sparse combinations of features representing human poses [2], [5], [18], [23], [26], [27]. Some methods estimated 3D human pose based on features like shape context [14], silhouettes [1], scale-invariant feature transform (SIFT) descriptors [4], and histogram of gradients (HOG) [21].

B. DEEP-BASED 3D HUMAN POSE ESTIMATION

With the availability of large collections like Human3.6M [9] and the effectiveness of deep learning, recently researchers develop deep learning methods to estimate 3D human pose. Some works [15], [16], [25] have been proposed to train a network in an end-to-end manner, i.e., input an image and estimate 3D human pose directly. Such methods often lead to limited results due to variations of brightness, color, or texture. Martinez *et al.* [13] divided the problem into two parts, i.e., detecting the 2D skeleton from the image first, and then inferring a 3D skeleton from the 2D skeleton. In this way, a simple linear model can be developed to achieve promising results. Our work follows this two-stage scheme.

Most previous works focused on estimating 3D human pose from a single frame. Recently, researchers have been trying to consider temporal information in the video to obtain more reliable predictions and reduce the influence of noise. Pavllo *et al.* [17] presented a simple and efficient method for 3D skeleton estimation in videos based on dilated temporal convolutions on a sequence of 2D skeletons.

In addition to using temporal information, if the action was captured from multiple views, information from different views may be complementary. Chen *et al.* [6] extracted features from two different views, and transformed the features from one view into another. If features from different views are transformed well, these features can be used to estimate better 3D human pose.

Supervised learning methods require a large amount of 2D images paired with 3D skeleton labels. However, labeling 3D data is laborious, and thus annotated 3D data are scarce. Therefore, some works [12], [17], [19] have been proposed to achieve 3D human pose estimation in a weakly supervised or semi-supervised way. Wandt *et al.* [22] proposed the idea of re-projection and adversarial learning, which enable the model to be effectively trained even if only weakly labeled 3D data are available.

In this work, we would like to jointly consider temporal information and multiview information mainly based on the framework proposed in [22], but train the framework based on the semi-supervised scheme. We develop this unified framework to take temporal and spatial factors together, and verify effectiveness of the proposed method.

III. PROPOSED METHOD

We first briefly introduce the reprojection network (RepNet) proposed in [22], and then present details of how to integrate multiple video frames and multiple views into a unified framework.

A. REPROJECTION NETWORK (RepNet)

The RepNet was designed to take 2D skeletons as the input, and focus on estimating the corresponding 3D skeletons in a weakly supervised manner. Figure 1 illustrates the idea

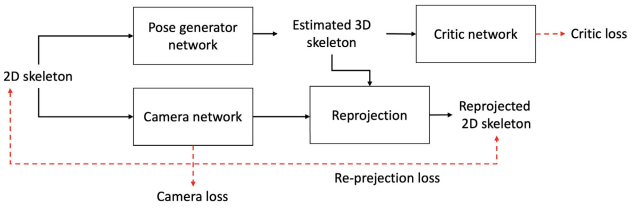

FIGURE 1. Illustration of the re-projection network proposed in [22].

of RepNet. Given a 2D skeleton $\boldsymbol{S} = (x_1, y_1, \ldots, x_n, y_n) \in \mathcal{R}^{2\times n}$, where x_i and y_i are the xy coordinates of the ith joint, a pose generator network G is developed to estimate the corresponding 3D skeleton $G(\boldsymbol{S}) = \boldsymbol{T}$. The 3D skeleton $\boldsymbol{T}$ is represented by $(x'_1, y'_1, z'_1, \ldots, x'_n, y'_n, z'_n) \in \mathcal{R}^{3\times n}$, where x'_i, y'_i, and z'_i are xyz coordinates of the ith joint. Conventional supervised methods compare the estimated 3D skeleton with the ground truth to guide training of the pose generator network. However, the 3D ground truth is scarce. To resolve this issue, three components were proposed in [22].

First, whether the estimated 3D skeleton is plausible is examined by a critic network C. This critic network is designed to measure the difference between the distribution of estimated 3D skeletons and the distribution of real 3D skeletons. To train this critic network, the Wasserstein loss function $\mathcal{L}_{crt}$ defined in [3] is used.

Second, the camera parameters $\boldsymbol{K}$ estimated by the camera network should describe weak perspective projection. In [22], this property is introduced and used to design the camera loss $\mathcal{L}_{cam}$, in order to guide camera network training. Conceptually, the loss is calculated as the Frobenius norm between the normalized $\boldsymbol{K}\boldsymbol{K}^T$ and identity projection [22].

Third, based on the estimated camera parameters $\boldsymbol{K}$, the estimated 3D skeleton $\boldsymbol{T}$ can be re-projected to a 2D one $\boldsymbol{S}' = \boldsymbol{K}\boldsymbol{T}$, which should be similar to the given input $\boldsymbol{S}$. To measure the difference, the Frobenius norm between $\boldsymbol{S}$ and $\boldsymbol{S}'$ is calculated as the reprojection loss $\mathcal{L}_{rep}$.

Overall, three losses $\mathcal{L}_{crt}$, $\mathcal{L}_{cam}$, and $\mathcal{L}_{rep}$ are linearly combined as $\mathcal{L}_{overall} = \mathcal{L}_{crt} + \mathcal{L}_{cam} + \mathcal{L}_{rep}$, in order to guide training of the entire network. Implementation details of the RepNet please refer to [22].

B. CONSIDERING TEMPORAL INFORMATION

Separately predicting 3D skeletons for each video frame usually causes jiggling results. Inspired by [17], we would like to jointly consider 2D skeletons $\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_M$ in M consecutive frames, and train a pose generator network G to predict the 3D skeleton $\boldsymbol{T}_{M/2}$ in the $\frac{M}{2}$-th frame.

Figure 2 shows architecture of the pose generator network and the camera network that jointly considers 2D skeletons at multiple frames. The building blocks are basically residual blocks [7] consisting of convolutional layers with leaky ReLU as the activation function. Particularly, the input consists of 2D keypoints in M consecutive video frames. The 2D skeleton in each frame is constituted by 16 keypoints, and can be represented as a 32-dimensional vector. Therefore, the 2D skeletons in M frames form a $M \times 32$ matrix. The convolutional layer denoted by $2J$, $3d1$, and 256, for example, means that the input channel is $2J$, the convolution kernel is 3×3 with dilation 1, and the output channel is 256.

The first half of Figure 2 extracts features from the given 2D skeletons. The second half is constituted of two branches, one for pose generation and the other for camera parameter estimation. The pose generator branch outputs a 48-dimensional result representing coordinates of the sixteen 3D keypoints in the $\frac{M}{2}$-th frame. The camera branch outputs a 6-dimensional result representing the camera parameters, based on which 3D skeletons can be reprojected back into 2D skeletons.

To guide model training, three losses $\mathcal{L}_{crt}$, $\mathcal{L}_{cam}$, and $\mathcal{L}_{rep}$ are calculated and linearly combined as the same as mentioned in Sec. III-A. We jointly consider information from $\boldsymbol{S}_1$ to $\boldsymbol{S}_M$, and we predict the 3D skeleton corresponding to the middle of the input sequence, i.e., $\boldsymbol{T}_{M/2}$ in the $\frac{M}{2}$-th frame. To calculate losses, the ground truths corresponding to the $\frac{M}{2}$-th frame are used. Training data are sliding windows of M consecutive 2D skeletons, with stride 1. When the frame we want to estimate is at the beginning or at the end of the video, we just pad with the skeleton of the first frame or the last frame, making the number of skeletons reach M.

C. CONSIDERING MULTIVIEW INFORMATION

The Human3.6M dataset was collected by simultaneously capturing the same individual's action from multiple viewpoints. Jointly considering multiple views enables us to include richer information for 3D pose estimation. In addition to temporal information, we attempt to include multiview information in the unified framework, as shown in Figure 3.

Taking two views as an example, the inputs are two 2D skeleton sequences $\boldsymbol{S}_1^{(1)}, \boldsymbol{S}_2^{(1)}, \ldots, \boldsymbol{S}_M^{(1)}$ and $\boldsymbol{S}_1^{(2)}, \boldsymbol{S}_2^{(2)}, \ldots, \boldsymbol{S}_M^{(2)}$, and the estimated 3D skeletons corresponding to the $\frac{M}{2}$-th frame are $\boldsymbol{T}_{M/2}^{(1)}$ and $\boldsymbol{T}_{M/2}^{(2)}$, respectively. The Wasserstein losses $\mathcal{L}_{crt}^{(1)}$ and $\mathcal{L}_{crt}^{(2)}$ are respectively calculated based on the

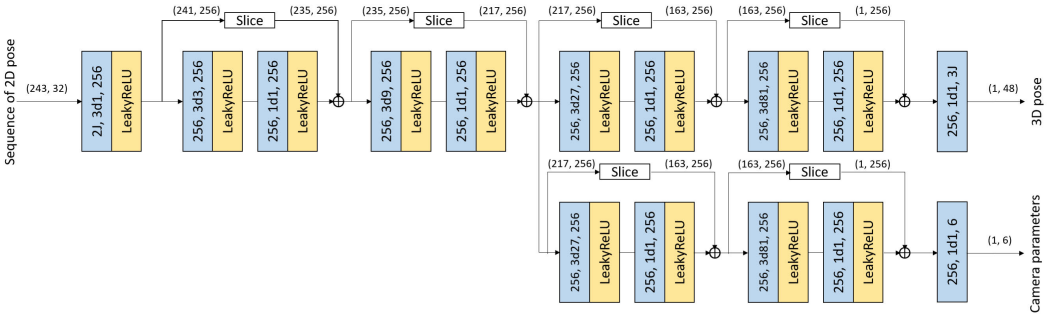

FIGURE 2. Architecture of the pose generator network and the camera network that jointly consider 2D skeletons in multiple frames.

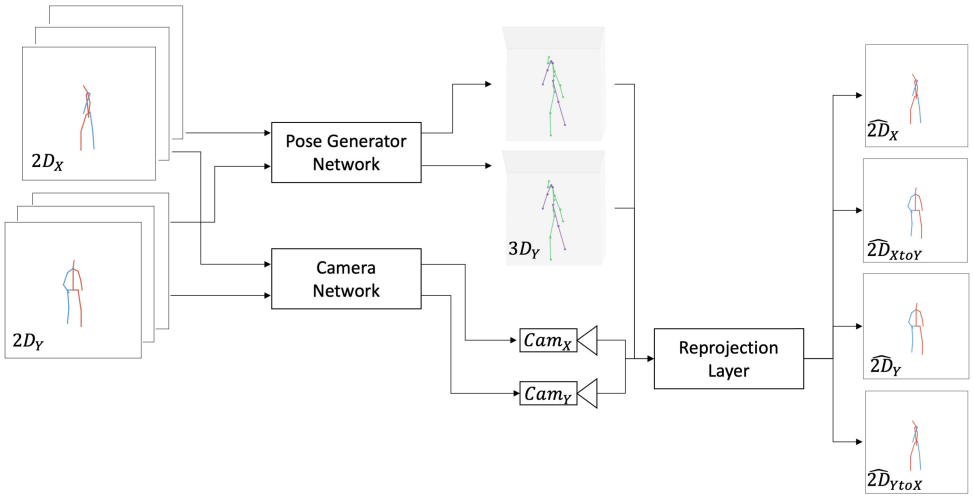

FIGURE 3. Architecture of the framework that jointly considers multiple viewpoints.

critic network C. Similarly, we respectively estimate camera parameters $\boldsymbol{K}^{(1)}$ and $\boldsymbol{K}^{(2)}$ for two viewpoints by the camera network, and calculate the camera losses $\mathcal{L}_{cam}^{(1)}$ and $\mathcal{L}_{cam}^{(2)}$, respectively.

One thing very important is that, for the 2D skeleton sequence $\boldsymbol{S}_i^{(1)}$ at the first viewpoint, the corresponding pose generator network estimates the 3D skeleton $\boldsymbol{T}_i^{(1)}$ at the predefined reference viewpoint V. For the 2D skeleton sequence $\boldsymbol{S}_i^{(2)}$ at second viewpoint, the corresponding pose generator network estimates the 3D skeleton $\boldsymbol{T}_i^{(2)}$ also at the reference viewpoint V. Therefore, no matter based on 2D skeletons from which viewpoints, the estimated 3D skeletons are at the same viewpoint. This can be seen from the middle part of Figure 3. The camera parameters $\boldsymbol{K}^{(1)}$ and $\boldsymbol{K}^{(2)}$ estimated by two camera networks thus can be used to project estimated 3D skeletons from viewpoint V to the first viewpoint and the second viewpoint, respectively.

By the camera parameters $\boldsymbol{K}^{(1)}$, we reproject $\boldsymbol{T}_{M/2}^{(1)}$ into $\boldsymbol{S}_{M/2}^{\prime(1)}$, and compare it with $\boldsymbol{S}_{M/2}^{(1)}$ to get reprojection loss $\mathcal{L}_{rep}^{(1)}$. Similarly, by $\boldsymbol{K}^{(2)}$, we reproject $\boldsymbol{T}_{M/2}^{(2)}$ into $\boldsymbol{S}_{M/2}^{\prime(2)}$, and compare it with $\boldsymbol{S}_{M/2}^{(2)}$ to get reprojection loss $\mathcal{L}_{rep}^{(2)}$. More interestingly, we can do cross-view reprojection and tighten the relationship bewteen two views. By $\boldsymbol{K}^{(1)}$, we project $\boldsymbol{T}_{M/2}^{(2)}$ into $\boldsymbol{S}_{M/2}^{\prime(2\to1)}$, which should be similar to $\boldsymbol{S}_{M/2}^{(1)}$ if the estimated 3D skeleton and the camera parameters are correct. We thus calculate the cross-view loss $\mathcal{L}_{rep}^{(2\to1)}$. Similarly, we can also calculate the cross-view loss $\mathcal{L}_{rep}^{(1\to2)}$ by projecting $\boldsymbol{T}_{M/2}^{(1)}$ into $\boldsymbol{S}_{M/2}^{\prime(1\to2)}$.

To combine the aforementioned losses, the basic idea is that we equally treat importance of two different views. We also equally treat four cases of the reprojection losses. Overall, the losses mentioned above are linearly combined as:

$$\begin{aligned}\mathcal{L}_{overall} &= (0.5\mathcal{L}_{crt}^{(1)} + 0.5\mathcal{L}_{crt}^{(2)}) + (0.5\mathcal{L}_{cam}^{(1)} + 0.5\mathcal{L}_{cam}^{(2)}) \\ &\quad + (0.25\mathcal{L}_{rep}^{(1)} + 0.25\mathcal{L}_{rep}^{(2)} + 0.25\mathcal{L}_{rep}^{(1\to2)} + 0.25\mathcal{L}_{rep}^{(2\to1)}).\end{aligned} \tag{1}$$

We also empirically tried different weights, but the experimental results are similar or worse.

IV. EVALUATION

A. DATASET AND EXPERIMENTAL SETTINGS

We mainly perform experiments on the Human3.6M dataset [9]. Videos in this dataset were acquired by recording 15 types of actions performed by 5 female and 6 male subjects, under 4 different viewpoints. In a well-set environment, 4 fixed-position digital video cameras were used to simultaneously capture the subject's action from 4 corners of a rectangular room, and 10 motion cameras were rigged on the walls to capture the signals from small reflective markers attached to the subject's body. By tracking and calibrating these signals, 3D coordinates of body joints are labeled. Of the 11 subjects, seven are annotated with 3D poses. Figure 4 shows some screenshots of videos in the Human3.6M dataset captured from four viewpoints. Overall, this dataset contains 3.6 million video frames for 11 subjects from four different viewpoints. According to the settings of many previous works [22], we evaluate on 17-joint human skeleton. The joint at the hip is always set as the origin. Therefore, the human pose estimation network predicts the coordinates of the remaining 16 joints relative to the hip.

Following the experimental settings in [19], we train and evaluate the proposed network in a semi-supervised manner. At the training stage, 2D skeletons of the five subjects S1, S5, S6, S7, and S8, and only the 3D skeletons of S1 are used. This design is to simulate that 2D labeled data are available, but much fewer 3D labeled data are available for training. At the test stage, 2D skeletons of the subjects S9 and S11 are taken as the inputs to estimate 3D skeletons.

The camera matrix contains rotational and scaling components. To avoid ambiguities between the camera and 3D pose rotation, all the rotational and scaling components from the 3D poses are removed. We align every 3D pose to a template

TABLE 1. Performance Comparison Between the Proposed Method and the State-of-the-Art Supervised Methods in Terms of MPJPE.

Methods	Direc.	Disc.	Eat	Greet.	Phone	Photo	Pose	Purch.
Martinez et al. [13]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1
Pavlakos et al. [16]	41.2	49.2	42.8	43.4	55.6	46.9	40.3	63.7
Pavllo et al. [17]	45.1	47.4	42.0	46.0	49.1	56.7	44.5	44.4
Chen et al. [6]	41.1	44.2	44.9	45.9	46.5	39.3	41.6	54.8
RepNet	62.3	88.4	80.4	73.7	79.3	98.0	71.8	81.3
RepNet+M	54.2	66.4	53.8	59.1	60.3	76.4	55.9	59.7
RepNet+T	59.0	83.9	80.6	71.5	82.8	100.6	65.0	82.7
RepNet+T+M	49.1	63.6	48.6	56.0	57.4	69.6	50.4	62.0
Methods	Sit	SitD	Smoke	Wait	Walk	WalkD	WalkT	Avg.
Martinez et al. [13]	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Pavlakos et al. [16]	97.6	119.9	52.1	42.7	51.9	41.8	39.4	56.9
Pavllo et al. [17]	57.2	66.1	47.5	44.8	49.2	32.6	34.0	47.1
Chen et al. [6]	73.2	46.2	48.7	42.1	35.8	46.6	38.5	46.3
RepNet	107.4	120.5	77.1	77.9	100.2	64.6	52.5	82.4
RepNet+M	80.2	80.2	61.9	56.6	63.6	45.8	45.9	61.3
RepNet+T	100.4	113.3	76.6	71.7	97.0	57.0	49.8	79.5
RepNet+T+M	75.4	77.4	57.2	53.5	57.7	37.6	38.1	56.9

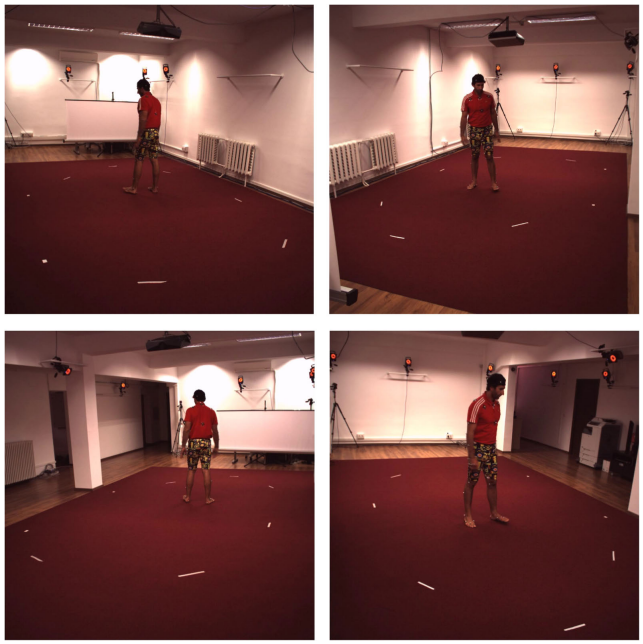

FIGURE 4. Screenshots of videos in the Human3.6M dataset captured from four viewpoints.

pose via the procrustes alignment [8], as shown in the middle part of Figure 3.

Two metrics are used to measure performance. The first one is the mean per-joint position error (MPJPE) in millimeters [9], which is the mean Euclidean distance between predicted joint positions and ground-truth joint positions. The second is the error after alignment with the ground truth in translation, rotation, and scale (P-MPJPE) [6], [13], [17].

B. COMPARING WITH SUPERVISED METHODS

We first compare the proposed semi-supervised method with other supervised methods. To train the supervised models, all 2D and 3D data of S1, S5, S6, S7, and S8 are used. Specifically, totally 1,688,984 (1.68 million) frames and thus 1.68 million 3D labeled skeletons are used in the supervised learning methods. Like other works of supervised learning, we use 2D data with a full frame rate of 50 fps. According to [17], when considering temporal information in the full frame rate setting, the length of input fragments is 243 frames.

Table 1 shows performance comparison between the proposed method and the state-of-the-art supervised methods, in terms of MPJPE values averaged over 15 different actions in the Human3.6M dataset. The baseline RepNet achieves 82.4 MPJPE, which is inferior to all supervised methods. This is not surprising. The RepNet was originally proposed to train in a weakly-supervised method, and the 3D ground truth is not explicitly compared with the estimated 3D skeleton. Based on the RepNet architecture, we obtain performance improvement when multiview information (denoted as "RepNet+M)" or temporal information (denoted as "RepNet+T)" is considered. Jointly considering temporal and multiview information significantly improves RepNet from 82.4 MPJPE to 56.9 MPJPE (based on the one-tail Student's t test, the p value is 8.8e-05). Comparing with the supervised methods shown in the first half of Table 1, we see that RepNet with the designed improvements can achieve encouraging results competitive with [13] and [16].

Figure 5 shows some sample results on the Human3.6M dataset when two types of information are jointly considered (RepNet+T+M).

C. COMPARING WITH SEMI-SUPERVISED METHODS

We next compare the proposed method with current semi-supervised methods. To train the semi-supervised model, 2D data of S1, S5, S6, S7, and S8, and only the 3D data of S1 are used for training. Specifically, only the 3D data in 271,436 frames from S1 are used as the 3D labels, which are much fewer than 1.68 million as mentioned in Sec. IV-B. The main goal of this comparison is to show that, even without most 3D labeled data, the semi-supervised model can still achieve competitive performance.

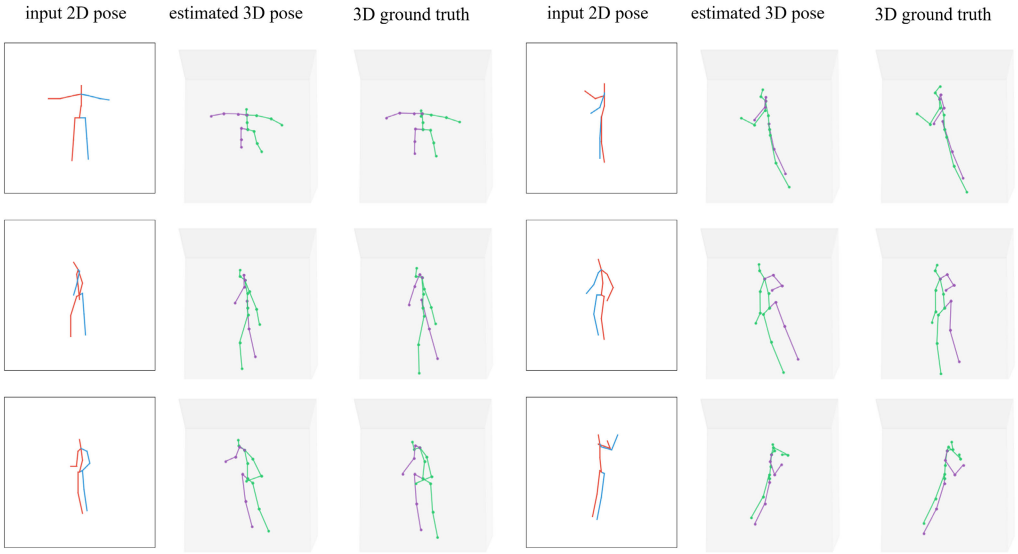

FIGURE 5. Sample results on the Human3.6M dataset.

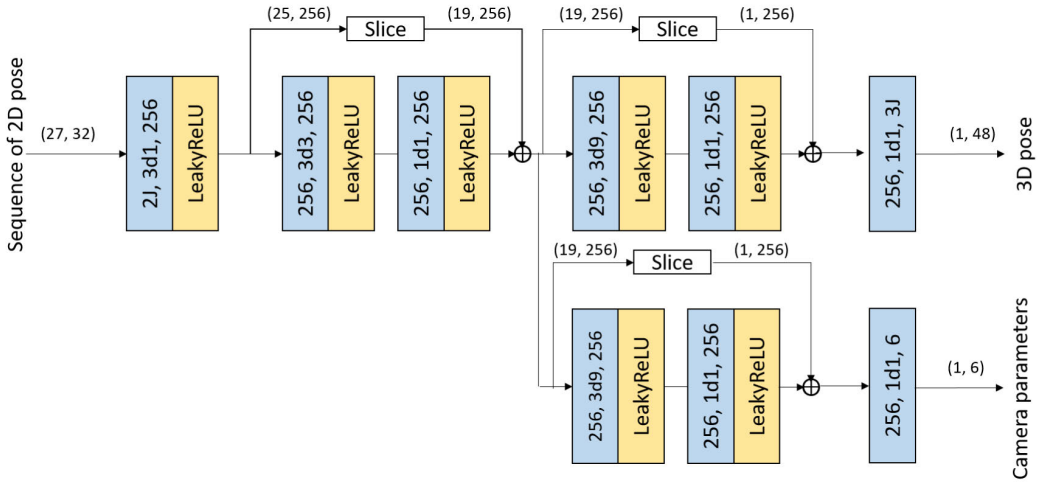

FIGURE 6. Architecture of the pose generator network and the camera network that can take 27 frames as the inputs.

At this stage of comparison, all 2D and 3D data are downsampled to 10 fps according to the setting mentioned in [19]. Because the data is downsampled, we modify the network architecture proposed in [17] and integrate it into the RepNet framework. The pose generator network and the camera network are designed to be able to take 27 frames as the inputs, as shown in Figure 6.

1) EFFECTIVENESS OF ADVERSARIAL LEARNING

The idea of weak supervision shown in Figure 1 is jointly accomplished by the critic network and the re-projection layer. One may think that re-projection has already enables weak supervision. To verify effectiveness of the critic network, we intentionally remove influence of the critic network by removing the loss $\mathcal{L}_{crt}$ when training, and see how performance changes.

Table 2 shows performance of the framework without the critic network, in terms of MPJPE. As can be seen, the performance becomes very poor without the critic network. This verifies necessity of the critic network and the effectiveness of adversarial learning.

2) COMPARING WITH SOTA

Table 3 shows performance comparison between the proposed method and the state-of-the-art semi-supervised methods, in terms of MPJPE. In our method, when we only consider multiview information, performance of "RepNet+M" outperforms the baseline RepNet. But if we

TABLE 2. Performance of the Proposed Framework Without the Critic Network, in Terms of MPJPE.

Methods	Average MPJPE
RepNet	588.9
RepNet+M	1168.7
RepNet+T	377.7
RepNet+T+M	755.0

TABLE 3. Performance Comparison Between Semi-Supervised Methods, in Terms of MPJPE.

Methods	Average MPJPE
Rhodin et al. [19] (multiview)	131.70
Pavllo et al. [17] (temporal)	85.0
Li et al. [12] (temporal)	88.80
RepNet	110.9
RepNet+M	92.6
RepNet+T	126.7
RepNet+T+M	86.7
RepNet+T+M+S1-supervised	**61.2**

only consider temporal information, the performance drops. It may be because the amount of data is decreased and the network structure appears too simple. Relatively it is more difficult to well train the generator and discriminator in this situation. When considering both temporal and multiview information "RepNet+T+M", the performance is better than considering one factor only.

In order to make full use of 3D data for training, we further make the following settings. We especially pick out the data belonging to S1 in each mini batch, calculate the mean square error between the estimated 3D pose and the corresponding ground truth, and use this loss to fine-tune the pose generator network once. The setting "RepNet+T+M+S1-supervised" shows that, with this finetuning, the obtained results largely outperform existing works.

Inputs of the proposed method and the RepNet are 2D skeletons. Quality of the input 2D skeletons thus obviously influences the final performance. To verify the influence, we purposely input ground truth 2D skeletons and estimate the corresponding 3D ones. Table 4 shows performance difference between approaches when truth 2D skeletons are taken as the inputs or not, in terms of MPJPE. The row "RepNet+T+M+S1-supervised (w. 2DGT)" shows that, by taking truth 2D skeletons as the inputs, the MPJPE value largely decreases from 61.2 to 54.8. This shows importance of quality of the input 2D skeletons.

TABLE 4. Performance Difference Between Approaches When Truth 2D Skeletons are Taken as the Inputs or not, in Terms of MPJPE.

Methods	Average MPJPE
RepNet+T+M+S1-supervised (w/o 2DGT)	61.2
RepNet+T+M+S1-supervised (w. 2DGT)	54.8

Table 5 shows performance comparison between state-of-the-art semi-supervised methods, in terms of P-MPJPE.

TABLE 5. Performance Comparison Between Semi-Supervised Methods, in Terms of P-MPJPE.

Methods	Average P-MPJPE
Rhodin et al. [19] (multiview)	98.2
Pavllo et al. [17] (temporal)	61.6
Li et al. [12] (temporal)	66.5
RepNet+T+M+S1-supervised (w/o 2DGT)	50.3
RepNet+T+M+S1-supervised (w. 2DGT)	45.6

Again we see the proposed method outperforms existing methods. The row "RepNet+T+M+S1-supervised (w. 2DGT)" shows that, by taking truth 2D skeletons as the inputs, the P-MPJPE value decreases from 50.3 to 45.6. Comparing with the difference shown in Table 4, the performance difference in terms of P-MPJPE is slightly smaller than MPJPE. We think this is contribution of the procrustes alignment.

V. CONCLUSION

We have presented a semi-supervised 3D human pose estimation framework that jointly takes temporally consecutive frames captured from multiple viewpoints. Based on a limited amount of 3D labeled data, this network predicts 3D skeletons from the given 2D skeletons, and estimate camera parameters as well. Based on the estimated camera parameters, the predicted 3D skeletons are re-projected into 2D ones, which should be similar to the input 2D skeletons. In this manner, a semi-supervised 3D pose estimation network is constructed. In this article, we largely enhance this network by jointly considering temporal information and multiview information. Through comprehensive evaluation, we show that the proposed network achieves the state-of-the-art performance, comparing to other semi-supervised methods.

Occlusion has always been one of most challenging problems in 3D human pose estimation. Currently we take temporally consecutive frames captured from multiple viewpoints as richer resources to extract more representative features. In the future, we would like to design occlusion-aware mechanisms and more completely take advantage of temporal information and multiview information. In addition, there may be difference between the same action taken by different individuals due to cultural difference or gender. We may be able to leverage embedding matching like [24] to learn individual-invariant representations.

REFERENCES

[1] A. Agarwal and B. Triggs, "3D human pose from silhouettes by relevance vector regression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2004, p. 2.
[2] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1446–1455.
[3] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.
[4] L. Bo, C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Fast algorithms for large scale conditional 3D prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[5] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 561–578.
[6] X. Chen, K.-Y. Lin, W. Liu, C. Qian, and L. Lin, "Weakly-supervised discovery of geometry-aware representation for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10895–10904.
[7] 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.
[8] J. R. Hurley and R. B. Cattell, "The procrustes program: Producing direct rotation to test a hypothesized factor structure," *Behav. Sci.*, vol. 7, no. 2, pp. 258–262, Jan. 2007.
[9] 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, Aug. 2014.
[10] H. Jiang, "3D human pose reconstruction using millions of exemplars," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 1674–1677.
[11] H. J. Lee and Z. Chen, "Determination of 3D human body postures from a single view.," *Comput. Vis., Graph. Image Process.*, vol. 30, no. 2, pp. 148–168, 1985.
[12] Z. Li, X. Wang, F. Wang, and P. Jiang, "On boosting single-frame 3D human pose estimation via monocular videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2192–2201.
[13] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2640–2649.
[14] G. Mori and J. Malik, "Recovering 3D human body configurations using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1052–1062, Jul. 2006.
[15] S. Park, J. Hwang, and N. Kwak, "3D human pose estimation using convolutional neural networks with 2D pose information," *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 156–169.
[16] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-Fine volumetric prediction for single-image 3D human pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7025–7034.
[17] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7753–7762.
[18] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Reconstructing 3D human pose from 2D image landmarks," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 573–586.
[19] 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.
[20] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net++: Multi-person 2D and 3D pose detection in natural images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1146–1161, May 2019.
[21] Shakhnarovich, Viola, and Darrell, "Fast pose estimation with parameter-sensitive hashing," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, p. 750.
[22] 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. 7782–7791.
[23] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao, "Robust estimation of 3D human poses from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2014, pp. 2361–2368.
[24] Z. Wang, B. Du, and Y. Guo, "Domain adaptation with neural embedding matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2387–2397, Jul. 2020.
[25] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei, "Deep kinematic pose regression," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 186–201.
[26] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis, "Sparse representation for 3D shape estimation: A convex relaxation approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1648–1661, Aug. 2017.
[27] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "Sparseness meets deepness: 3D human pose estimation from monocular video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4966–4975.

WEI-TA CHU (Senior Member, IEEE) received the B.S. and M.S. degrees in computer science from National Chi Nan University, Taiwan, in 2000 and 2002, respectively, and the Ph.D. degree in computer science from National Taiwan University, Taiwan, in 2006. From 2007 to 2019, he was a Professor with National Chung Cheng University. From July 2008 to August 2008, he was also a Visiting Scholar with the Digital Video and Multimedia Laboratory, Columbia University. From January 2017 to March 2017, he was also a Visiting Professor with Nagoya University. He is currently a Professor with the Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan. His advised master's students received several thesis awards from the Taiwan Institute of Electrical and Electronic Engineering, Institute of Information and Computing Machinery, and the Chinese Institute of Electrical Engineering. His research interests include digital content analysis, multimedia indexing, deep learning, and pattern recognition.

Dr. Chu was a recipient of the Best Full Technical Paper Award in ACM Multimedia 2006. He was also a recipient of the Young Faculty Awards presented by National Chung Cheng University, in 2011, the K. T. Li Young Researcher Award presented by Institute of Information and Computing Machinery, in 2012, the Best GOLD Member Award presented by the IEEE Tainan Section, in 2013, the Distinguished Alumni Award presented by National Chi Nan University, in 2014, and the Outstanding Youth Electrical Engineer Award by the Chinese Institute of Electrical Engineering, in 2017. He serves as a Program Co-Chair for ACM ICMR 2021, MMM 2020, and IEEE MMSP 2019. He is an Associate Editor of *IEICE Transactions on Information and Systems*, from 2016 to 2020.

ZONG-WEI PAN received the bachelor's degree from the Department of Mathematics, National Chung Cheng University, in 2018, and the master's degree from the Department of Computer Science and Information Engineering, National Chung Cheng University, in 2020. His research interests include image processing, machine learning, and multimedia analysis.