HAL
open science

Selective Spatio-Temporal Aggregation Based Pose Refinement System: Towards Understanding Human Activities in Real-World Videos

Di Yang, Rui Dai, Yaohui Wang, Rupayan Mallick, Luca Minciullo, Gianpiero Francesca, Francois F Bremond

To cite this version:

Di Yang, Rui Dai, Yaohui Wang, Rupayan Mallick, Luca Minciullo, et al.. Selective Spatio-Temporal Aggregation Based Pose Refinement System: Towards Understanding Human Activities in Real-World Videos. WACV 2021 – IEEE Winter Conference on Applications of Computer Vision, Jan 2021, Virtual, United States. hal-03121883

HAL Id: hal-03121883

https://hal.science/hal-03121883v1

Submitted on 26 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Selective Spatio-Temporal Aggregation Based Pose Refinement System: Towards Understanding Human Activities in Real-World Videos

Di Yang[1,2] Rui Dai[1,2] Yaohui Wang[1,2]
Rupayan Mallick[1] Luca Minciullo[3] Gianpiero Francesca[3] François Brémond[1,2]

[1]Inria [2]Université Côte d'Azur [3]Toyota Motor Europe
{di.yang, rui.dai, yaohui.wang, rupayan.mallick, francois.bremond}@inria.fr
{luca.minciullo, gianpiero.francesca}@toyota-europe.com

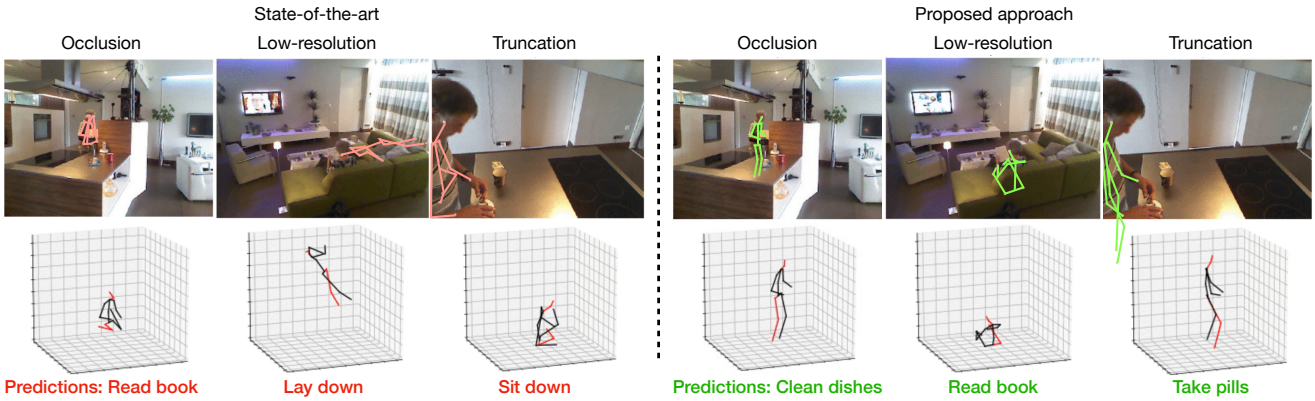

Figure 1. **Skeleton-based action recognition** on Toyota Smarthome with poses (Left) extracted by AlphaPose [8] (left), LCRNet++ [30] (middle), OpenPose [2] (right) and high-quality poses (Right) obtained by the proposed pose refinement system. The action predictions using refined poses with the same action recognition system become more accurate. (3D reconstructions are from VideoPose [24] over 2D)

Abstract

Taking advantage of human pose data for understanding human activities has attracted much attention these days. However, state-of-the-art pose estimators struggle in obtaining high-quality 2D or 3D pose data due to occlusion, truncation and low-resolution in real-world un-annotated videos. Hence, in this work, we propose 1) a Selective Spatio-Temporal Aggregation mechanism, named SST-A, that refines and smooths the keypoint locations extracted by multiple expert pose estimators, 2) an effective weakly-supervised self-training framework which leverages the aggregated poses as pseudo ground-truth instead of handcrafted annotations for real-world pose estimation. Extensive experiments are conducted for evaluating not only the upstream pose refinement but also the downstream action recognition performance on four datasets, Toyota Smarthome, NTU-RGB+D, Charades, and Kinetics-50. We demonstrate that the skeleton data refined by our Pose-Refinement system (SSTA-PRS) is effective at boosting various existing action recognition models, which achieves competitive or state-of-the-art performance. Refined pose data is available at: https://github.com/walker-ally/SSTA-PRS

1. Introduction

Action recognition approaches have led to significant improvements for many applications, such as video surveillance, video understanding, human-computer interaction and game control. Compared with RGB-based action recognition using spatio-temporal deep convolution on RGB videos [37, 3, 44, 31], skeleton-based approaches have drawn increasing attention owing to their strong ability in summarizing human motion [40, 48, 39, 50, 45, 17, 1, 46, 18, 35, 9, 34, 33, 19]. As high level representations, skeletons have the merits of being robust to changes of appearances, environments, and view-points. Earlier skeleton-based approaches using deep-learning like RNNs [48, 39, 50, 38, 45] or temporal CNNs [17, 1] are proposed owing to their high representation capacity, but they ignore the important semantic connectivity of the human body. Recent GCNs-based approaches [46, 18, 35, 9, 34, 33, 19] construct

spatial-temporal graphs and model the spatial relationships within GCNs directly, and these methods have seen significant performance boost, indicating the interest of using semantic human skeleton for action recognition. However, most of these algorithms use accurate 3D human poses obtained from a Motion Capture system [32] and they cannot perform well on real-world videos with low-quality 2D or 3D poses. In the real-world, 1) the performance of human pose estimation approaches [30, 12, 2, 8, 15, 41, 11] is limited, especially in case of occlusion, truncation, and low-resolution scenarios. 2) A large amount of manual skeleton annotation is extremely expensive to obtain. Motivated by the above problem, we propose a skeleton-based action recognition framework that uses a Selective Spatio-Temporal Aggregation based Pose Refinement System, named SSTA-PRS, to **extract high-quality 2D skeletons from un-annotated real-world videos** and leverages AGCNs [35] for action classification and action detection.

To deal with **the absence of keypoints (*i.e.* joints) due to occlusion, truncation and low-resolution** in real-world pose estimation (as shown in Fig. 1), we construct a multi-expert pose estimation system to predict improved 2D poses. It is fine-tuned with the pseudo ground-truth 2D pose generated by a novel Selective Spatio-Temporal Aggregation (SST-A) which integrates the pose proposals computed from several existing expert pose estimators. In this work, we select LCRNet++ [30], OpenPose [2] and AlphaPose [8] as the experts. Regarding on our downstream tasks, we leverage the AGCNs [35] to extract features for an input pose sequence that are used directly for action classification (for trimmed videos) or fed into a temporal model like LSTM [10], TCNs [16] for temporal action detection (for untrimmed videos).

The contributions related to this paper are summarized as follows: 1) We propose a novel Selective Spatio-Temporal Aggregation mechanism (SST-A), that integrates the advantage of several expert pose estimation systems in both spatial and temporal domains, and we introduce a confidence metric C to evaluate the quality of the aggregated poses. 2) We present a weakly-supervised self-training Pose Refinement System (SSAT-PRS) based on LCRNet++ [30] using pseudo-ground truth poses, generated by our SST-A mechanism instead of using handcrafted pose annotations. 3) We directly evaluate the upstream pose refinement method using pose ground-truth then demonstrate the effectiveness of this framework for the downstream GCNs-based action recognition task using action ground-truth. Experiments are conducted on four real-world datasets, Toyota Smarthome, NTU-Pose, Charades and Kinetics-50.

2. Related Work

Human Pose Estimation in Real-World. Most state-of-the-art approaches for 2D human pose estimation employ 2D CNNs architectures for a single image in a strongly-supervised setting [23, 2, 41, 8, 12, 4, 15, 20]. For 3D pose estimation, [30, 22] focus on end-to-end reconstruction by directly estimating 3D poses from RGB images without intermediate supervision. [49] applies GCNs for regression tasks, especially 2D to 3D human pose regression. [24] demonstrates that 3D poses in video can be effectively estimated with a fully convolutional model based on dilated TCNs over 2D keypoint sequences. Among these methods, [30, 23, 22, 8, 12, 41] have first to incorporate a person detector, followed by the estimation of the joints and then the computation of the pose for each person. These approaches give full-body predictions once the people are detected, but the detection speed slows down with the increase of the number of people present in the image. [2, 4, 15, 20] are bottom-up approaches which detect all joints in the image using heatmaps that estimate the probability of each pixel to correspond to a particular joint, followed by associating body parts belonging to distinct individuals. These approaches cannot always provide the none-visible body parts for each individual due to occlusions and truncations.

By annotating poses in the real-world, approaches [30, 2, 8, 29] are becoming more robust to occlusion and they can provide us with pre-trained pose estimators, so that we can extract skeleton data from real-world videos without expensive handcraft annotations. In particular, LCRNet++ [30] is an attractive pose estimator, which leverages a Faster R-CNN [28] like architecture with a CNNs backbone. A Region Proposal Network extracts candidate boxes around humans. To deal with occlusions and truncation, LCRNet++ proposes 'anchor-poses' for pose classes instead of object classes: these key poses typically correspond to a person standing, sitting, etc. Bottom-up method OpenPose [2] proposes an alternative approach by regressing affinities between joints (*i.e.* the direction of the bones), together with the heatmaps. AlphaPose [8] improves the performance of top-down pose estimation algorithms by detecting accurate human poses even with inaccurate bounding boxes. Closer to our work, Rockwell et al. [29] propose an effective self-training framework that adapts human 3D mesh recovery systems to consumer videos. They focus on recovering from occlusions and truncations, but they do not have solutions to tackle low-resolution images and the instability of the extracted 3D meshes along time. In our work, we combine the advantages of the three expert pose estimators [30, 2, 8] by spatio-temporal aggregating their results and getting a more accurate pose than using only one expert.

Skeleton-Based Action Classification. ST-GCN [46], where spatial graph convolutions along with interleaving temporal convolutions are used for spatial-temporal modeling is the first GCNs-based method for action recognition. Subsequently, many GCNs methods representing a more appropriate spatial graph on intra-frames have been proposed. 2s-AGCN [35] introduces an adaptive graph convolutional network to adaptively learn the topology of the graph with self-attention, which can better suit the action

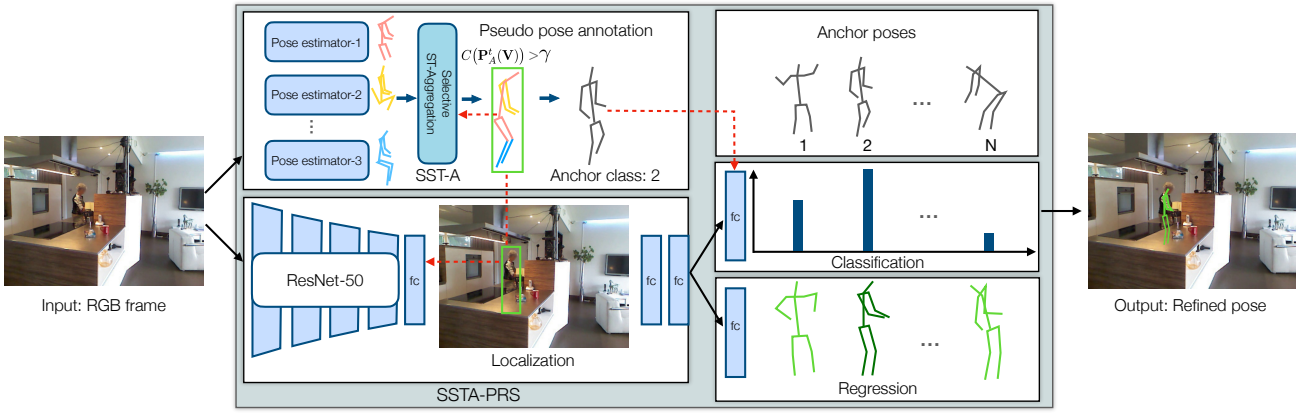

Figure 2. **Overview of Pose-Refinement System (SSTA-PRS).** Given an RGB frame, a less noisy 2D pose $\mathbf{P}_A^t(\mathbf{V})$ and its confidence value C is computed by the Selective Spatio-Temporal Aggregation (SST-A) with the pose proposals obtained by several pose estimation systems and the previous aggregated pose. If the confidence is higher than the threshold γ, we are able to calculate the pseudo ground-truth bounding box and anchor class according to this improved pose to fine-tune the localization and classification branches of an LCRNet [30] architecture in a weakly-supervised setting. Finally, this refined pose estimation system is used to extract high-quality 2D poses in the real-world videos for the downstream action recognition task.

recognition task and the hierarchical structure of the GCNs. It also uses a two-stream ensemble with skeleton bone features to boost performance. Their further work, MS-AAGCN [34] proposes multi-stream adaptive graph convolutional networks that introduce attention modules and a multi-stream ensemble based on 2s-AGCN [35]. Note that these approaches primarily focus on spatial modeling. In contrast, MS-G3D [19] presents a unified approach for capturing complex joint correlations directly across spacetime.

Yan et al. [46] use OpenPose [2] to extract poses for applying GCNs on the real-world dataset [3], but the pose data is not always full-body due to occlusions and truncations.

Skeleton-Based Action Detection. Action detection (*i.e.* action localization) task needs to find precise temporal boundaries of actions occurring in an untrimmed video. In order to model long temporal relationships, the current detection methods [16, 26, 25] encode the videos as a preprocessing step. After that, action detection can be seen as a sequence-to-sequence problem. Recurrent neural networks (RNNs) [14, 47] have been popularly used to model actions transitions between frames. These approaches relied on the memory cell to accumulate temporal information in different frames, thus implicitly capturing the relationships between actions. Temporal convolution networks (TCNs) are another type of action detection model. Dilated-TCN [16] increases the temporal reception field by using dilated convolutions to model long temporal patterns. Temporal Gaussian Mixture [26] utilizes the Gaussian distribution to generate the weights of the temporal kernel. This Gaussian-based kernel enables TGM to capture long temporal relations via a large receptive field with limited parameters.

In our work, we apply AGCNs [35] that performs the best on Smarthome [6] with estimated and refined pose data instead of using a Motion Capture system. To the best of our knowledge, we are the first to provide high-quality full-body poses using a pose refinement system and leverage AGCNs [35] for both action classification and detection in real-world datasets [6, 36] without ground-truth poses.

3. Pose Refinement Approach

The overall architecture of the proposed method for pose refinement is shown in Fig. 2. Given an RGB frame, several pose proposals are obtained by multiple expert pose estimation systems [30, 12, 2, 8, 15, 41, 11] and the Selective Spatio-Temporal Aggregation mechanism (SST-A) computes an improved pose, which is more accurate, smoother, and more stable along time. With this aggregated pose, we compute a confidence metric to estimate its quality. Then, we select the aggregated poses with higher confidences than a threshold and calculate the pseudo ground-truth bounding box and anchor class to fine-tune the localization and classification branches of an LCRNet [30] architecture. Finally, the refined pose estimation system is used to extract higher-quality poses in real-world videos.

3.1. Selective Spatio-Temporal Aggregation

Selective Spatio-Temporal Aggregation (SST-A) is the key component to deal with the **absence of keypoints** caused by occlusion, truncation and low-resolution, and with the **instability in time domain** due to pose estimation from a single frame. Our insight is that 1) bottom-up methods directly predict the keypoints through the heatmaps, however, they may miss joints that are none-visible due to occlusions or truncations because the number of each body part prediction may not correspond to the number of people in the image. 2) Top-down methods regress the coordinates of keypoints over the bounding box of the people. As long as people are detected, the keypoints of full-body

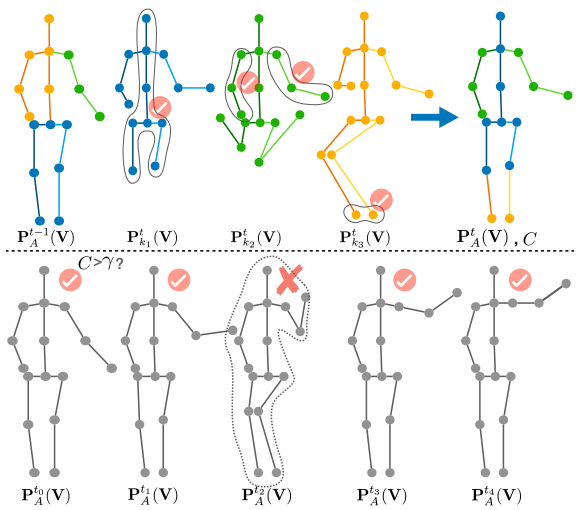

Figure 3. **Two steps of SST-Aggregation**. 1) Aggregation in both spatial and temporal level (top). The pose $\mathbf{P}_A^t(\mathbf{V})$ of current frame is aggregated from the three pose proposals $\mathbf{P}_{k_1}^t(\mathbf{V})$ (blue), $\mathbf{P}_{k_2}^t(\mathbf{V})$ (green) and $\mathbf{P}_{k_3}^t(\mathbf{V})$ (yellow). 2) Selective temporal filter (bottom). The pose with a low confidence in the aggregated sequence will be discarded.

can be predicted. But these methods may miss people in low-resolution images, resulting in missing all the joints of these people. According to the above analysis, by combining the results of both families of methods, we can reduce the number of missing joints and obtain more stable and higher-quality full-body keypoints. Therefore, we leverage multiple expert pose estimation systems, including methods from both families to extract poses for the same frame as several pose proposals, and then aggregate them to recover the missing keypoints. In this work, we select two top-down estimation systems LCRNet++ [30] and AlphaPose [8] and a bottom-up estimation system OpenPose [2] to provide the pose proposals, which are then combined into an improved pose sequence through our SST-A mechanism. Moreover, our pose sequence is extracted frame by frame with the estimators, so there is a certain lack of temporal continuity, resulting in a static joint shaking in the video. This problem is also an obstacle for the performance of action recognition. Hence, our SST-A also uses a temporal filtering mechanism to smooth the entire sequence by eliminating unstable values.

As shown in Fig. 3, we note all the N keypoints in one body as a set $\mathbf{V} = \{\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_N}\}$, the frames as a set $F = \{0, 1, 2, ..., T\}$ and the position of one joint $\mathbf{v}$ ($\mathbf{v} \in \mathbf{V}$) in the frame t ($t \in F$) estimated by the pose estimation system k_m ($k_m \in K = \{k_1, k_2, ..., k_M\}$) as $\mathbf{P}_{k_m}^t(\mathbf{v})$, noted that K is the ensemble of pose estimation systems. The final aggregated pose sequence of the body $\mathbf{V}$ is noted as $\mathbf{P}_A^F(\mathbf{V}) = \{\mathbf{P}_A^t(\mathbf{v}) | \mathbf{v} \in \mathbf{V}, t \in F\}$ and our aggregation system has two steps.

1) Joint-level aggregation. Each keypoint of the pose is calculated from the prediction results of the three estimators [2, 30, 8] in the current frame $\mathbf{P}_{k_m}^t(\mathbf{v})$ and the aggregated result of the previous frame $\mathbf{P}_A^{t-1}(\mathbf{v})$. So this step is to select the closest keypoint to the same part aggregated in the previous frame. For the first frame, we select any one of the two closest keypoints obtained by pose estimators. Joint-level aggregation can be written as:

$$\mathbf{P}_A^t(\mathbf{v}) = \begin{cases} \mathbf{P}_{k_a}^t(\mathbf{v}), \text{ if } t = 0 \\ (k_a, k_b) = \underset{(k_i,k_j)\in K^2, i\neq j}{\operatorname{argmin}} \Big(D\big(\mathbf{P}_{k_i}^t(\mathbf{v}), \mathbf{P}_{k_j}^t(\mathbf{v})\big) \Big) \\ \mathbf{P}_{k_a}^t(\mathbf{v}), \text{ if } t > 0 \\ k_a = \underset{k_i \in K}{\operatorname{argmin}} \Big(D\big(\mathbf{P}_{k_i}^t(\mathbf{v}), \mathbf{P}_A^{t-1}(\mathbf{v})\big) \Big) \end{cases} \tag{1}$$

where D is the Euclidean distance between two key-points in the image, noted as:

$$D\big(\mathbf{P}_1(\mathbf{v}), \mathbf{P}_2(\mathbf{v})\big) = \sqrt{\big(\mathbf{P}_1(\mathbf{v}) - \mathbf{P}_2(\mathbf{v})\big)^2} \tag{2}$$

2) Body-level aggregation. Followed by the first step which can effectively solve the problem of missing keypoints, we define a **confidence metric** $C \in (0, 1]$ that describes the likelihood that the aggregated pose is the real pose in order to further smooth the pose sequence. We believe that when the average similarity between the aggregated pose and the pose proposals is very high, the pose proposals are also very similar, indicating that the pose proposal itself is likely to be accurate, and the aggregation result will have a high confidence. This selective likelihood filter is written as (3), which is to discard the abnormal poses with a very low confidence in the whole sequence,

$$\mathbf{P}_A^t(\mathbf{V}) = \begin{cases} \mathbf{P}_A^t(\mathbf{V}), \text{ if } C\big(\mathbf{P}_A^t(\mathbf{V})\big) >= \gamma \\ \text{discard}, \text{ if } C\big(\mathbf{P}_A^t(\mathbf{V})\big) < \gamma \end{cases} \tag{3}$$

where C (4) is defined to describe the confidence of this aggregated pose. (D_{normal} is the distance between the aggregated head and neck while offset $\epsilon = 10^{-12}$ is to prevent errors in case of $D_{normal} = 0$)

$$C\big(\mathbf{P}_A^t(\mathbf{V})\big) = \exp\Big(-\frac{1}{NM} \sum_{\mathbf{V}} \sum_{K} \frac{D\big(\mathbf{P}_A^t(\mathbf{v}), \mathbf{P}_{k_m}^t(\mathbf{v})\big)}{D_{normal} + \epsilon} \Big) \tag{4}$$

γ is a filtering parameter that represents a threshold. If the confidence of the pose in the current frame is lower than this threshold, it will be discarded from the sequence. After this two-step SST-A is completed, we obtain a higher-quality full-body skeleton sequence from a video, which can be effectively used as pseudo ground-truth pose for our self-training Pose Refinement system in Sec. 3.2.

3.2. Self-Training Pose Refinement System

SST-A can effectively integrate the advantages of [30, 2, 8]. However, this aggregation method may increase the workload in practice because we have to estimate the poses several times with different systems. Hence, we propose a self-training framework using the higher-quality 2D poses

obtained from SST-A as supervised pseudo ground-truth, to refine one of the pose estimation models. Once the model is refined, the other models are not needed for inference. In fact, we only need to run SST-A on a small part of the dataset, and then it can be used for fine-tuning the network. As shown in Fig. 2, we build our pose refinement model (SSTA-PRS) based on LCRNet++ [30] owing to its particularity of its three branches (localization, classification and regression), we do not have to provide truly accurate pose labels but only fine-tune the localization and classification branches with the pseudo ground-truth 2D poses.

3.2.1 Overview of SSTA-PRS architecture

As LCRNet++ [30], our SSTA-PRS framework also contains 4 main components. 1) **Localization:** it leverages a Faster R-CNN [28] like architecture with a ResNet-50 backbone [13]. Given an input image, a Region Proposal Network (RPN) [28] extracts candidate boxes around humans. 2) **Classification:** these regions are then classified into different 'anchor-poses' pre-defined by K-means clustering that typically correspond to a person standing, a person sitting, etc. In this paper, 'anchor-poses' are defined in 2D only, and the refinement occurs in this joint 2D pose space. 3) **Regression:** a class-specific regression is applied to estimate body joints in 2D. First, for each class of pose, we define offline the 'anchor-poses', computed as the center over all elements in the corresponding cluster. After fitting all the 2D anchor-poses into each of the candidate boxes, we perform class-specific regressions to deform these anchor-poses and match the actual 2D pose in each box. 4) **Post-processing:** for each individual, multiple pose candidates can overlap and produce valid predictions. These pose candidates are combined by pose proposal integration [30], taking into account their 2D overlap and classification scores. As the approach is holistic, it outputs full-body poses, even in case of occlusions or truncation by image boundaries.

3.2.2 Weakly-supervised training

We train this model with a weakly-supervised setting, which only refines the 2D localization and classification. The reason is that firstly, our pseudo pose annotations are not sufficiently accurate for regression while they are accurate enough for localization and classification. Secondly, the in-the-wild pre-trained model has good prediction performance when the localization and classification are correct. However, in low-resolution images, the bounding boxes of people are usually very difficult to search, which may result in no estimated keypoint on the body, or an error in the classification stage leading to inaccurate pose prediction. Therefore, if the classification and localization branches are correctly fine-tuned, the model should find the correct anchor class so that the final prediction can be more accurate.

Pseudo 2D pose ground-truth: it contains two parts, the bounding box of people and the anchor class. Both are calculated using the SST-A pose results. We take the maximum and minimum values of the pose in x and y directions as the boundary of the initial bounding box. We then expand the box by 10% as ground-truth for the localization branch, because the key-points do not correspond exactly to the boundary of the person. The class label of pose P, noted as $Class_P \in \{0, 1, ..., B\}$, is set by finding the closest 2D anchor-pose $Anchor_P$ according to the similarity S [30] between the oriented 2D poses centered at the left-top corner of bounding box: $Class_P = \text{argmin}_b\, S(Anchor_b, P)$. This label is used by the classification branch as pseudo ground-truth.

Loss function: our loss is the sum of the following two losses, described as:

$$L = L_{loc} + L_{classif} \tag{5}$$

The loss of the localization component is the loss of the region proposal network [28] (RPN):

$$L_{loc} = L_{RPN} \tag{6}$$

Same as [30], let u be the probability distribution estimated by SSTA-PRS, obtained by the fully connected layers of the classification branch after RoI pooling, followed by a softmax function. The classification loss is defined using the standard cross entropy loss:

$$L_{classif}(u, Class_P) = -\log u(Class_P) \tag{7}$$

4. Experimental Setup

Our objective is to obtain high-quality poses from real-world videos in order to understand human activities. We conduct a wealth of experiments to evaluate our system with two protocols: 1) Evaluation by the upstream pose refinement task using pose ground-truth, which is to directly compare the accuracy of the poses obtained from the pose estimators [30, 2, 8] with our proposed SSTA-PRS. 2) Evaluation by the downstream action recognition task using ground-truth action labels. We use the same action recognition model [35], but processing different pose data with and without refinement for comparing the action recognition (classification and detection) performances, which indirectly demonstrates the effectiveness of our SSTA-PRS.

4.1. Datasets and Evaluation Protocols

Toyota Smarthome (Smarthome) [6] is a recent real-world daily living dataset for action classification, recorded in an apartment where 18 older subjects carry out tasks of daily living during a day. The dataset contains 16,115 videos of 31 action classes, and the videos are taken from 7 different camera viewpoints. All the actions are performed in a natural way without strong prior instructions. This dataset provides RGB data and 3D skeletons which are extracted from LCRNet++ [30]. In contrast, in our experiments, we extract higher-quality 2D skeleton data with

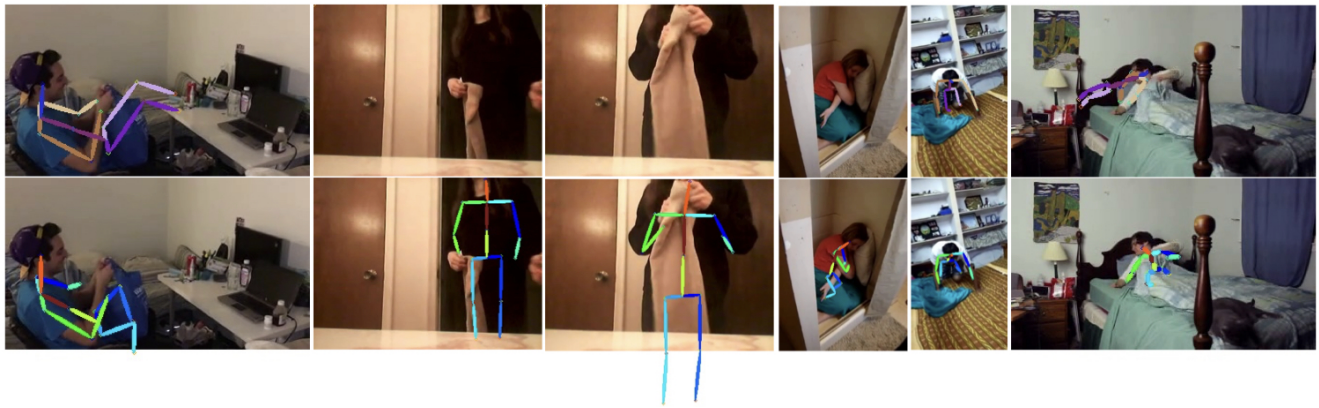

Figure 4. Comparison of the poses extracted on Charades by LCRNet++ [30] without refinement (top) with our SSTA-PRS (bottom).

our SSTA-PRS framework instead of the provided ones. For the evaluation on this dataset, we follow cross-subject (CS) and cross-view (CV1 and CV2) protocols. We use Smarthome as our main dataset for ablations and additional experiments[1].

Smarthome-Pose: in order to evaluate directly the poses extracted by our SSTA-PRS, we chose the middle frames for randomly selected 1,400 videos of Toyota Smarthome and we annotated the 2D poses to create a test set containing 1,400 images with 640×480 resolution and many occlusions, truncations. We follow the PCKh @0.5 (percent of keypoints within a threshold of 0.5 times head length) as the pose evaluation protocol. We regard the distance in pixels between head and neck as the head length.

NTU-Pose: NTU-RGB+D [32] is a large-scale multimodal dataset which consists of 56,880 sequences of high-quality 2D/3D skeletons with 25 joints, associated with depth maps, RGB and IR frames captured by the Microsoft Kinect v2 sensor. We selected 60 videos (6,098 frames) with the same subject performing different actions and we took the 2D skeleton as the ground-truth for pose evaluation. The dataset was recorded in a laboratory, so in this work, we changed the original quality of the videos by reducing the resolution to 320×180 and adding partial occlusions to make it similar to our real-world settings. We use the 2D MPJPE (mean per joint position error normalized by head length) and PCKh @2.0 protocols.

Charades [36] is a challenging dataset due to large environmental diversity. The videos were recorded in the homes of 267 subjects. While performing the actions, the subjects are usually occluded by objects or the furniture of the scene. We apply SSTA-PRS on Charades to validate whether this algorithm can extract robust skeletons even when subjects are occluded. (Fig. 4) Note that there are 157 action classes in the Charades dataset and most of them are relevant to objects but not to poses. As we only use the skeletons to recognize actions, we kept the 31 meaningful semantic verbs as the action classes. These semantic verbs and additional results using the original class labels are provided in supplementary material. We follow the localization setting of the dataset, reporting the frame-based mAP.

Kinetics-50 is a subset of a large-scale real-world dataset Kinetics-400 [3] that contains about 300,000 video clips for 400 action classes collected from YouTube and does not contain pose information. We took the first 50 action categories in alphabetical order corresponding to 27,314 video clips and we extracted the 2D poses with our SSTA-PRS module. Following the evaluation method in [46], we train the action recognition model on the training set (25,288 video clips) and report the top-1 and top-5 accuracy on the validation set (2,026 video clips).

4.2. Implementation Details

Pose estimators: we select 1) OpenPose 18-joints [2], 2) AlphaPose Sample-Baseline [8] using YOLOv3 [27] as detector and 3) LCRNet++ In-The-Wild [30] as three expert pose estimation models and LCRNet++ In-The-Wild [30] as the student model for refining the pose. The poses contain the 13 main joints that all three estimators can detect.

Pose refinement: we select all the videos from NTU-Pose, 10% of the videos from Smarthome and $9.0K$ videos from Charades, and 10% of the frames for each video with uniform sampling to get a large dataset containing $3.0K$ images from NTU-Pose, $40.2K$ images from Smarthome, and $65.9K$ from Charades. We then split 20% of the images as the validation set. We apply SST-A mechanism using [30, 8, 2] as k_1, k_2 and k_3 and take 13 main keypoints for aggregation with $\gamma = 0.18$ as the confidence threshold for temporal filter to generate pseudo ground-truth 2D poses (Sec. 3.1). Then, we use In-The-Wild pre-trained model of LCRNet++ [30], which sets 20 'anchor poses' and leverages ResNet50 [13] as a backbone and we follow the standard setting values from [30]. We fine-tune this model using 4 images per batch, and 512 boxes per image. The refined model is used to estimate 2D poses of the whole set of Smarthome, Charades and also for Kinetics-50 although its videos are not in the training set.

[1] Details of additional 5C-AGCN is in the supplementary material

5. Evaluation Using Pose Ground-Truth

To evaluate the performance of our upstream pose refinement system (SSTA-PRS), we experiment on NTU-Pose and Smarthome-Pose to directly compare the performance of expert pose estimators with our SSTA-PRS.

5.1. Results And Discussion

SST-A: We estimate 2D poses using three expert estimators and then perform SST-A without discarding any frames. The results in Tab. 1 show that SST-A is effective to integrate the advantages of the expert estimators and achieves a better performance ($+7.7\%$ on NTU-Pose, $+1.3\%$ on Smarthome-Pose).

Confidence metric: To analyze the reliability of the confidence metric (Sec. 3.1) that filters the poses for pseudo annotations, we analyse on the NTU-Pose the variation of the MPJPE with the confidence C and find that the confidence decreases globally with an increase of the error. We select $\gamma = 0.18$ as the confidence threshold that can keep most of the aggregated poses within the error of 3.0 (Visualization in the supplementary material). According to this threshold, we analyze the frequency of the retained (*i.e.* with C-high confidence) and discarded (*i.e.* with C-low confidence) poses within different error intervals (Fig. 5). Within the intervals of smaller errors, we keep the most of the poses and remove the ones in the larger error intervals. In order to have sufficient training samples, we still keep some poses with a few errors (but high confidence), corresponding to cases of complex scenes. Our fine-tuning system is weakly-supervised training, these poses can still play a positive role in localization and classification. Therefore, the confidence metric is instrumental in our work.

SSTA-PRS: After this filtering stage, we can get higher-quality pseudo 2D pose annotations for fine-tuning SSTA-PRS. Compared with the three expert estimators (Tab. 1), our fine-tuned SSTA-PRS is the most effective ($+13.9\%$ on NTU-Pose and $+9.3\%$ on Smarthome-Pose).

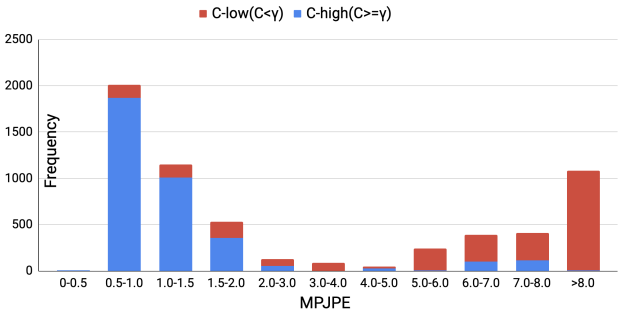

Figure 5. Histogram of pose frequency in function of MPJPE with threshold $\gamma = 0.18$ (*i.e.* high confidence when $C > \gamma$).

6. Evaluation Using Action Ground-Truth

We conduct several experiments on Toyota Smarthome, Charades and Kinetics-50 for the downstream action recognition task to evaluate the poses extracted through the proposed SSTA-PRS. We perform an ablation study to validate the effectiveness of our pose refinement system. Then, we compare our system with other state-of-the-art methods.

Methods	NTU-Pose PCKh @2.0 (%)	Smarthome-Pose PCKh @0.5 (%)
LCRNet++ [30]	54.1	64.4
AlphaPose [8]	53.2	55.5
OpenPose [2]	45.4	58.9
SST-A only(ours)	61.8	65.7
SSTA-PRS(ours)	**68.0**	**73.7**

Table 1. PCKh of poses from different pose estimators and proposed SSTA-PRS using SST-A only (Sec. 3.1) and using both SST-A and self-training (Sec. 3.2) on NTU-Pose and Smarthome-Pose.

6.1. Action Recognition Approach

We compare the performance of the best state-of-the-art action recognition model on Smarthome, AGCN [35], with and without our pose refinement module.

Action Classification: Unless stated, we only use 2D poses for action recognition, which reduces the computation costs without losing performance as we have accurate enough 2D poses. We call the implemented model 2s-AGCN+SSTA-PRS. We also expand the channels of 2s-AGCN [35] to five as 5C-AGCN1, which concatenates the centered 3D poses to handle relative motion dynamics and 2D poses to obtain the global trajectory information.

Action Detection: Similar to other action detection models [10, 26, 16], the input to the baseline models is the segment-level video encoding. In this work, we utilize AGCNs [35] to encode the skeleton stream. The model is fine-tuned on the training set. To extract the features, a video is divided into T non-overlapping segments, each segment consisting of 16 continuous poses. These segments are sent to the fine-tuned AGCNs model to extract the segment-level features. This video representation is the input sequence to the action detection models [10, 26, 16].

6.2. Ablation Study

We analyze the impact of our two components for pose refinement. Performance is reported as mean per-class accuracy on Smarthome using only joint data.

Impact of pose refinement: we verify the effectiveness of the proposed pose refinement system using SST-A only or both SST-A and self-training (full). The results on Smarthome [6] are shown in Tab. 2. We test the 2s-AGCN [35] with the original 2D and 3D pose data extracted from LCRNet++ [30], and 2D pose refined by our SSTA-PRS including or excluding the self-training step. Note that the 3D refined pose data is obtained by VideoPose3D [24] over the corresponding 2D poses. It shows that both components can improve the performance on action classification ($+1 \sim 2\%$ brought by SST-A and $+2 \sim 5\%$ by self-training). In addition, we show in Tab. 3 the results for two training modes, with either weak or strong supervision (i.e.

with or without joint positions). We find that the results under strong supervision are below the ones using SST-A only. It suggests that the refined joint positions are too noisy to improve the regression branch. It shows also that SST-A significantly improves the upstream localization and classification performance of SSTA-PRS, which is sufficient to improve action classification.

2s-AGCN-Joint [35]	**Smarthome**	
	CS(%)	CV2(%)
+2D-original [6]	52.9	47.5
+2D SSTA-PRS (SST-A only)	53.5	47.9
+2D SSTA-PRS (full)	**55.7**	**49.0**
+3D-original [6]	49.1	49.6
+3D SSTA-PRS (SST-A only)	51.2	50.6
+3D SSTA-PRS (full)	**54.0**	**53.3**

Table 2. Mean per-class accuracy on Smarthome dataset using 2s-AGCN-Joint [35] with the skeleton data obtained by LCR-Net++ [30] (original), SSTA-PRS using SST-A only (Sec. 3.1) and using both SST-A and self-training(full) (Sec. 3.2).

SSTA-PRS Training Methods	**Smarthome** CS(%)
Strong-supervision	51.4
Weak-supervision	**54.0**

Table 3. Mean per-class accuracy on Smarthome dataset using the 3D skeleton data obtained by SSTA-PRS (2D) and VideoPose [24] (3D reconstruction over 2D skeleton)

6.3. Comparison With State-of-the-art

Toyota Smarthome: we compare the final model, 2s-AGCN+SSTA-PRS, with the state-of-the-art methods on the Toyota Smarthome dataset. The results shown in Tab. 4 demonstrate that high-quality pose is a good modality for real-world action classification and our method achieves state-of-the-art performance with a large margin ($+3.8\%$ on CS and on CV2) over the Pose-based methods. Moreover, we apply our module for VPN [7] as VPN+SSTA-PRS to show that the refined poses can also improve RGB-and-pose methods ($+4.4\%$ on CS and $+0.6\%$ on CV2).

Charades: we are the first to use only skeleton data on Charades, so we only compare the frame-based mAP of the same baseline methods using different poses as input to validate the effectiveness of the pose estimation algorithm. In this work, we use three types of input: 1) LCRNet3D poses extracted from LCRNet++ [30], 2) LCR-Net2D+VideoPose3D which uses video-based pose estimation system VideoPose3D [24] to reconstruct the LCR-Net2D poses to 3D poses, 3) SSTA-PRS 2D poses refined by our proposed algorithm. We utilize three baseline methods to evaluate these poses: Bidirectional-LSTM [10], Dilated-TCN [16] and TGM [26]. As shown in Tab. 5, SSTA-PRS is consistently better than image-based LCR-Net3D poses and video-based LCRNet2D+VideoPose3D poses for these three baselines (up to $+1.7\%$, $+0.6\%$ w.r.t. LCRNet3D and LCRNet2D+VideoPose3D respectively).

Kinetics-50: we compare the final model with the state-of-the-art skeleton-based action recognition methods on the Kinetics-50 dataset (Tab. 6). The methods used for comparison are the graph-based methods [35, 34] using OpenPose [2] skeleton data extracted by Yan et al. [46]. Our model based on SSTA-PRS refined poses achieves state-of-the-art performance (44.7% on Top1 and 73.1% on Top5).

Methods	**RGB**	**Pose**	**Smarthome**		
			CS(%)	CV1(%)	CV2(%)
DT [42]	✓	×	41.9	20.9	23.7
I3D [3]	✓	×	53.4	34.9	45.1
I3D+NL [43]	✓	×	53.6	34.3	43.9
AssembleNet++ [31](+object)	✓	×	**63.6**	-	-
P-I3D [5]	✓	✓	54.2	35.1	50.3
Separable STA [6]	✓	✓	54.2	35.2	50.3
VPN [7]	✓	✓	60.8	**43.8**	53.5
VPN+SSTA-PRS(ours)	✓	✓	**65.2**	**-**	**54.1**
LSTM [21]	×	✓	42.5	13.4	17.2
MS-AAGCN [34]	×	✓	56.5	-	-
2s-AGCN [35]	×	✓	57.1	22.1	49.7
2s-AGCN+SSTA-PRS(ours)	×	✓	**60.9**	**22.5**	**53.5**
5C-AGCN+SSTA-PRS(ours)1	×	✓	**62.1**	**22.8**	**54.0**

Table 4. Mean per-class accuracy comparison against state-of-the-art methods on the Toyota Smarthome dataset.

Methods	**Charades**		
	Bi-LSTM [10]	Dilated-TCN [16]	TGM [26]
LCRNet 3D [30]	17.6	18.0	18.9
LCRNet 2D+VideoPose [24]	18.9	18.7	20.1
SSTA-PRS (ours)	**19.5**	**19.5**	**20.6**

Table 5. Action detection performance on Charades using frame-based mAP.

Methods	**Kinetics-50**	
	Top-1(%)	Top-5(%)
2s-AGCN [35]	40.0	71.8
MS-AAGCN [34]	41.0	72.4
2s-AGCN+SSTA-PRS (ours)	**44.7**	**73.1**

Table 6. Classification accuracy comparison against state-of-the-art methods on the Kinetics-50 dataset.

7. Conclusion

In this work, we propose a novel method to extract pose sequences from challenging real-world videos. Owing to the proposed novel aggregation mechanism (SST-A) and weakly-supervised self-training framework, our method can be applied on 1) videos in low-resolution and 2) videos containing human body occlusions and truncations. When applied on a real-world activity recognition dataset (*e.g.* Smarthome), the poses extracted from our model can be used to achieve the best results which show the effectiveness of our method. However, it still has limitations since we mainly focus on improving the quality of poses. For the tasks concerning human-object interactions (*e.g.* Charades), our method does not bring too much boosting. One of the future directions could be combining our method with RGB-based methods together to further improve the recognition results.

Acknowledgement This work is supported by Toyota Motor Europe (TME). We are grateful to Sophia Antipolis - Mediterranean "NEF" computation cluster for providing resources and support.

References

[1] C. Caetano, F. Brémond, and W. Schwartz. Skeleton image representation for 3d action recognition based on tree structure and reference joints. *SIBGRAPI*, 2019.

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE TPAMI*, 2019.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[4] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020.

[5] S. Das, Arpit Chaudhary, F. Bremond, and M. Thonnat. Where to focus on for human action recognition? *WACV*, 2019.

[6] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *ICCV*, 2019.

[7] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. *ECCV*, 2020.

[8] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[9] X. Gao, W. Hu, Jiaxiang Tang, J. Liu, and Zongming Guo. Optimized skeleton-based action recognition via sparsified graph regression. *ACM MM*, 2019.

[10] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *IJCNN*, 2005.

[11] Riza Alp Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. *CVPR*, 2018.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[15] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *CVPR*, 2019.

[16] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, 2017.

[17] Chao Li, Qiaoyong Zhong, Di Xie, and S. Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *IJCAI*, 2018.

[18] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019.

[19] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, 2020.

[20] D. Luvizon, D. Picard, and H. Tabia. Multi-task deep learning for real-time 3d human pose estimation and action recognition. *IEEE TPAMI*, 2020.

[21] Behrooz Mahasseni and S. Todorovic. Regularizing long short term memory with 3d human-skeleton sequences for action recognition. *CVPR*, 2016.

[22] Gyeongsik Moon, Juyong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *ICCV*, 2019.

[23] G. Ning, Z. Zhang, and Z. He. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE TMM*, 2017.

[24] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019.

[25] AJ Piergiovanni and Michael S Ryoo. Learning latent super-events to detect multiple activities in videos. In *CVPR*, 2018.

[26] AJ Piergiovanni and Michael S. Ryoo. Temporal gaussian mixture layer for videos. In *ICML*, 2019.

[27] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *ArXiv*, 2018.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[29] Chris Rockwell and David F. Fouhey. Full-body awareness from partial observations. *ECCV*, 2020.

[30] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *IEEE TPAMI*, 2019.

[31] M. Ryoo, A. Piergiovanni, Juhana Kangaspunta, and A. Angelova. Assemblenet++: Assembling modality representations via attention connections. *ECCV*, 2020.

[32] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. *CVPR*, 2016.

[33] L. Shi, Yifan Zhang, Jian Cheng, and H. Lu. Skeleton-based action recognition with directed graph neural networks. *CVPR*, 2019.

[34] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing LU. Skeleton-Based Action Recognition with Multi-Stream Adaptive Graph Convolutional Networks. *arXiv:1912.06971 [cs]*, 2019.

[35] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019.

[36] Gunnar A. Sigurdsson, Gül Varol, X. Wang, Ali Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016.

[37] K. Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014.

[38] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, 2017.

[39] Amor Ben Tanfous, Hassen Drira, and Boulbaba Ben Amor. Sparse coding of shape trajectories for facial expression and action recognition. *IEEE TPAMI*, 2019.

[40] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. *CVPR*, 2014.

[41] Haoqian Wang, W. P. An, Xingzheng Wang, Lu Fang, and Jiahui Yuan. Magnify-net for multi-person 2d pose estimation. *ICME*, 2018.

[42] H. Wang, Alexander Kläser, C. Schmid, and C. Liu. Action recognition by dense trajectories. *CVPR*, 2011.

[43] X. Wang, Ross B. Girshick, A. Gupta, and Kaiming He. Non-local neural networks. *CVPR*, 2018.

[44] Y. Wang, L. Zhou, and Yu Qiao. Temporal hallucinating for action recognition with few still images. *CVPR*, 2018.

[45] Chunyu Xie, Ce Li, B. Zhang, Chen Chen, Jungong Han, Changqing Zou, and Jianzhuang Liu. Memory attention networks for skeleton-based action recognition. *IJCAI*, 2018.

[46] S. Yan, Yuanjun Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI*, 2018.

[47] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *IJCV*, 2018.

[48] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE TPAMI*, 2019.

[49] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *CVPR*, 2019.

[50] W. Zheng, L. Li, Z. Zhang, Y. Huang, and L. Wang. Relational network for skeleton-based action recognition. In *ICME*, 2019.