VPN++: Rethinking Video-Pose embeddings for understanding Activities of Daily Living

Srijan Das, Rui Dai, Di Yang, Francois F Bremond

VPN++: Rethinking Video-Pose embeddings for understanding Activities of Daily Living

Srijan Das, Rui Dai, Di Yang, Francois Bremond

Abstract—Many attempts have been made towards combining RGB and 3D poses for the recognition of Activities of Daily Living (ADL). ADL may look very similar and often necessitate to model fine-grained details to distinguish them. Because the recent 3D ConvNets are too rigid to capture the subtle visual patterns across an action, this research direction is dominated by methods combining RGB and 3D Poses. But the cost of computing 3D poses from RGB stream is high in the absence of appropriate sensors. This limits the usage of aforementioned approaches in real-world applications requiring low latency. Then, how to best take advantage of 3D Poses for recognizing ADL?

To this end, we propose an extension of a pose driven attention mechanism: Video-Pose Network (VPN), exploring two distinct directions. One is to transfer the Pose knowledge into RGB through a feature-level distillation and the other towards mimicking pose driven attention through an attention-level distillation. Finally, these two approaches are integrated into a single model, we call **VPN++**. We show that VPN++ is not only effective but also provides a high speed up and high resilience to noisy Poses. VPN++, with or without 3D Poses, outperforms the representative baselines on 4 public datasets. Code is available at https://github.com/srijandas07/vpnplusplus.

Index Terms—trimmed videos, pose, activities of daily living, embedding, attention.

1 Introduction

Learning representations for human actions, taking into account only the RGB modality is not sufficient. As a consequence, a large corpus of research studies has been focusing on multi-modal action recognition. The most popular and effective method is the two-stream approach [1], [2], [3] where one stream models appearance by taking RGB frames and the other stream models short-term motion by taking optical flow frames. However, this method is effective on videos obtained from web [4], [5], [6] where the human actions have prominent motion patterns. But what about Activities of Daily Living (ADL) where actions have subtle motion and often pertain to have similar spatio-temporal patterns?

Activities of Daily Living (ADL) may look simple but their recognition is often more challenging than activities present in sport, movie or Youtube videos. ADL often have very low inter-class variance making the task of discriminating them from one another very challenging. The challenges characterizing ADL are illustrated in fig 1: (i) short and subtle actions like *pouring water* and *pouring grain* while *making coffee* ; (ii) actions exhibiting similar visual patterns while differing in motion patterns like *rubbing hands* and *clapping*; and finally, (iii) actions observed from different camera views. In the recent literature, the main focus is the recognition of actions from internet videos [3], [9], [10], [11], [12] and very few studies have attempted to recognize ADL in indoor scenarios [13], [14], [15]. For instance, state-of-the-art 3D convolutional networks like I3D [3] pre-trained on huge video datasets [4], [5], [6] have successfully boosted the recognition of actions from internet videos. But, these networks with similar spatio-temporal kernels applied across the whole space-time volume cannot address the complex challenges exhibited by ADL. Attention mechanisms have thus been proposed on top of these 3D convolutional networks to guide them along the regions of interest of the targeted actions [9], [12], [13].

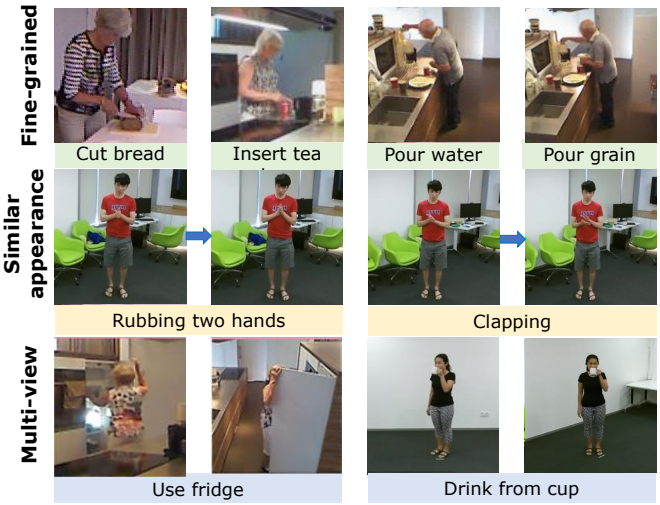

Fig. 1: Illustration of the challenges in Activities of Daily Living: fine-grained actions (top), actions with similar visual pattern (middle) and actions viewed from different cameras (below).

- *S. Das is with Stony Brook University, 100 Nicolls Rd, Stony Brook, NY 11794, USA.*
  E-mail: srijan.das@stonybrook.edu
- *R. Dai, D. Yang, and F. Bremond are with the Inria and Universite Cote d'Azur, 2004 Route des Lucioles, 06902 Valbonne, France.*
  E-mail: {rui.dai, di.yang, francois.bremond}@inria.fr

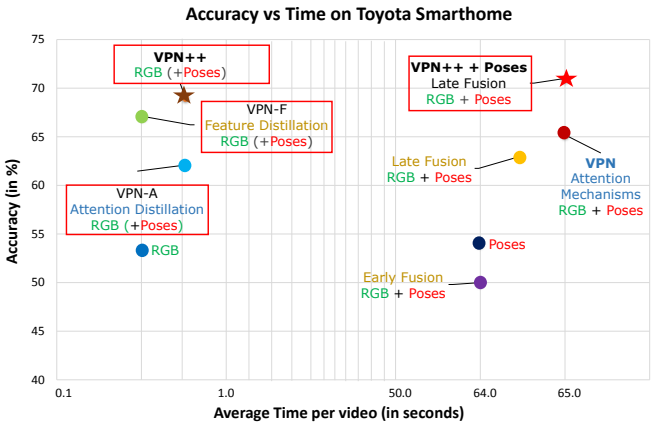

Fig. 2: Accuracy vs Time plot on Toyota Smarthome dataset for RGB and Pose modalities. 3D Poses are estimated using LCRNet++ [7] followed by Videopose3D [8]. Early fusion indicates concatenation of features at the last layer before prediction whereas Late fusion indicates averaging the prediction from both modalities. Our proposed models (marked with bounding box): **VPN-F**, **VPN-A** and **VPN++** mimicking Pose stream, outperforms all other RGB and Pose combining strategies, while being significantly faster. Late fusion of the distilled models with Pose stream further boosts the classification accuracy, but at the price of the model efficiency. Note that the model with input modalities denoted by RGB (+Poses) have been trained with RGB and Poses but do not require Poses at inference time.

Towards another approach, recent studies [15], [16], [17] have shown that human 3D poses provide a strong clue for understanding human-centric patterns in videos. Of course the use of 3D poses for human action analysis depends on (i) the availability of good quality 3D poses and (ii) architectures processing them. Thanks to algorithms like LCRNet++ [7] and VideoPose3D [8], high quality 3D poses can be obtained from RGB without the requirement of depth sensors. Similarly, the advancement of Graph based CNN architectures [18], [19], [20], [21] that take into account the human joint configurations have greatly impacted the skeleton based action recognition. Since skeleton based action recognition does not leverage the appearance information in videos, combining 3D poses and RGB is the need of the hour as studied in [14], [22], [23], [24], [25], [26], [27], [28].

The most common strategy for combining RGB stream and 3D poses includes (i) feature or score level fusion [22], [23], [24], [25]. As these modalities are heterogeneous, they must be processed by different kinds of network to show their effectiveness. This limits their performance in simple multi-modal fusion strategy [22], [23], [29]. Therefore, another approach adopted in recent days includes (ii) pose driven attention mechanisms [14], [26], [27], [28]. However, these methods have improved the action recognition performance but they do not take into account the alignment of the RGB cues and the corresponding 3D poses. Therefore, we proposed a spatial embedding to project the visual features and the 3D poses in the same referential in [30].

Further, this embedding is accompanied by an attention network to recognize a large variety of human actions. Thus, VPN consists of a spatial embedding and an attention network. It exhibits the following properties through its modules: (i) a spatial embedding learns an accurate video-pose embedding to enforce the relationships between the visual content and 3D poses, (ii) an attention network learns the attention weights with a tight spatio-temporal coupling for better modulating the RGB feature map, (iii) the attention network takes the spatial layout of the human body into account by processing the 3D poses through Graph Convolutional Networks (GCNs).

VPN to some extent overcomes the challenge of combining two modalities that are not only semantically different but also processed through heterogeneous networks. To go beyond these approaches, we study novel manners to combine efficiently RGB and 3D Poses. In particular, we aim at relaxing the need of high quality 3D poses, which are not always available. In Figure 2, we provide a plot of action classification accuracy vs average inference time on Toyota Smarthome [15] dataset. From the plot, we observe that feature level fusion (Early Fusion) performs worse since such fusion mechanisms are often prone to over-fitting [31] owing to an increase in the number of parameters of the network. Besides, Pose driven attention mechanism [30] yields high classification accuracy compared to RGB [3] and Poses [20] individually or their score level fusion (Late Fusion). But these RGB+Poses based methods are significantly slower than the RGB ones.

To this end, we explore the concept of knowledge distillation to infuse pose stream into RGB stream. Towards this objective, we propose two levels of distillation - one taking an approach of feature level fusion and the other one benefiting from attention mechanism. First, we aim at transferring feature-level knowledge from Pose to RGB stream to learn discriminative representation for recognizing actions, we call this feature-level distillation model **VPN-F**. To learn VPN-F, we use contrastive learning for distilling the knowledge from Pose stream to RGB. Besides avoiding the computation of poses at inference time, VPN-F learns to maximize the salient information from both streams towards action recognition. Second, we mimic pose driven attention network as in VPN through RGB stream. This is performed by adding a self-attention block in the RGB stream that hallucinates attention weights learned through 3D poses for the task of action recognition. We call this attention-level distillation model **VPN-A**. As an end result, VPN-A learns to provide pose driven attention weights which not only improve the action classification accuracy but also eliminate the requirement of poses at inference time. Finally, we integrate both levels of distillation into a single model called **VPN++**. Our experiments confirm that VPN++ is 160 times faster than the state-of-the-art methods without compromising effectiveness in real-world scenarios as illustrated in fig. 2. We also show that VPN++ via distillation when combined with 3D Poses, if available, outperforms the state-of-the-art results on 4 public datasets. Thus, to sum up, by infusing Poses into RGB using distillation, we provide a choice of highly effective models to the community that can be leveraged based on their needs like low latency, low sensitivity towards noisy Poses, or none.

2 Related Work

Significant improvement has been made in the action recognition domain after the advancement of 3D CNN [32]. Carreira and Zisserman [3] proposed a 3D CNN based fully convolutional network namely I3D which is pre-trained on huge datasets like Kinetics [5] to capture discriminative spatio-temporal patterns within an action. With the success of I3D, holistic methods like Pseudo 3D CNN [33], Separable 3D CNN [34], slow-fast network [10], channel-separated CNN [35], and X3D [36] have been fabricated for generic video datasets like Kinetics [5], UCF-101 [4] and HMDB [6]. But these networks with similar kernels applied across the whole space-time volume of a video, are too rigid to capture salient features for subtle patterns in ADL. Recently several attention mechanisms have been proposed on top of the aforementioned 3D ConvNets to extract salient spatio-temporal patterns. For instance, Wang et al. [9] have proposed a non-local module on top of I3D which computes the attention of each pixel as a weighted sum of the features of all pixels in the space-time volume. But this module relies too much on the appearance of the actions, i.e., pixel position within the space-time volume. As a consequence, this module though effective for the classification of actions in internet videos, fails to disambiguate ADL with similar motion and fails to address view-invariant challenges.

On the other hand, temporal evolution of 3D poses has been leveraged through sequential networks like LSTM and GRU for skeleton based action recognition [37], [38], [39]. Taking a step ahead, LSTMs have also been used for spatial and temporal attention mechanisms to focus on the salient human joints and key temporal frames [40]. Another framework represents 3D poses as pseudo images to leverage the successful image classification CNNs for action classification [41], [42]. Moreover, skeleton based action recognition has made significant improvements with the advancement of Graph Convolutional Networks (GCNs) [18], [19], [20], [21]. The key idea is to feed a graph representation of a skeleton frame in these networks which are optimized for the task of action classification. These graph based methods make use of the spatial topology of the human body joints and thus are more effective than recurrent networks [16], [38]. However, the skeleton based action recognition lacks in encoding the appearance information which is critical for ADL, such as in human-object interactions.

Combining modalities: Combining the advantages of privileged modalities in order to make use of their complementary discriminative power has been exploited widely in action recognition domain. Two-stream architectures [1], [2], [3] that learn separate features from optical flow and RGB modalities, outperform single modality approaches. Towards this direction, Ryoo et al. [43], [44] have proposed a Neural Search Architecture (NAS) to combine both RGB and Optical flow streams. In contrast to these methods, two complementary strategies are adopted to combine RGB and pose modalities. One is fusion of both modalities in feature space [22], [23], [24], [25]. However, these modalities are heterogeneous and must be processed by different kinds of network to show their effectiveness. Combining these heterogeneous features from different modalities through feature/score fusion introduce noise resulting in a downgraded action recognition performance [45]. The second is pose driven attention mechanisms to guide the RGB cues for action recognition as in [15], [26], [27], [28]. In [14], [26], [27], the pose driven attention networks implemented through LSTMs, focus on the salient image features and the key frames. Then, with the success of 3D CNNs, 3D poses have been exploited to compute the attention weights of a spatio-temporal feature map. Das et al. [28] have proposed a spatial attention mechanism on top of 3D ConvNets to weight the pertinent human body parts relevant for an action. Then, authors in [15] have proposed a more general spatial and temporal attention mechanism in a dissociated manner. But these methods have the following drawbacks: (i) there is no accurate correspondence between the 3D poses and the RGB cues in the process of computing the attention weights [14], [15], [26], [27], [28]; (ii) the attention sub-networks [14], [15], [26], [27], [28] neglect the topology of the human body while computing the attention weights; (iii) the attention weights in [15], [28] provide identical spatial attention along the video. As a result, action pairs with similar appearance like *jumping* and *hopping* are mis-classified. Therefore in [30], we proposed a new spatial embedding to enforce the correspondences between RGB and 3D poses which has been missing in the state-of-the-art methods. The embedding is built upon an end-to-end learnable attention network. The attention network considers the human topology to better activate the relevant body joints for computing the attention weights. To the best of our knowledge, none of the previous action recognition methods have combined human topology with RGB cues. In addition, the proposed attention network couples the spatial and temporal attention weights in order to provide spatial attention weights varying along time.

However, all the above approaches including VPN rely on the availability of 3D Poses, which not only escalates the model inference time but also increases their sensitivity towards Pose quality. Therefore, we use the concept of distillation that not only learns discriminative video-pose representations for understanding actions but also relaxes the demand for Poses at inference time. Consequently, we adopt both strategies to enforce RGB stream to (i) mimic pose stream features, and (ii) emulate pose-driven attention mechanism.

Distillation: Many approaches have exploited the concept of distillation for cross-modal knowledge transfer [29], [46], [47], [48], [49], [50], [51]. Towards action recognition, Garcia et al. [49] proposed a distillation framework consisting of teacher-student networks that hallucinates depth features from RGB features. This distillation is performed via logits as well as by matching feature maps of RGB and depth networks. Similarly, distillation approaches dynamically leveraging complementary information across several modalities have been proposed in [29], [51]. Crasto et al. [50] proposed MARS to train a RGB stream with standard cross-entropy loss along with mimicking the features learned by an optical flow stream. This mimicking is accomplished by a distillation loss that minimizes the euclidean distance between the learned features across both streams.

Thus, many distillation methods have been studied in the action recognition domain with OF and RGB, but not with RGB and Poses. Infusing Poses into RGB stream through distillation is not straightforward and includes

two main challenges: (i) 2D RGB images with appearance information and 3D Poses with geometric details are fed to the teacher-student heterogeneous networks, limiting the knowledge transfer between them due to their asymmetric dimensionality; (ii) the teacher network, i.e. the Pose stream is not consistently effective on the entire data distribution. In fact, the Pose stream carries irrelevant features for actions that are discriminated using their appearance information. Therefore, we propose to minimize the distance between the features learned by RGB & Poses while learning discriminative representation in the RGB feature space. Towards another approach with an effective teacher network, we perform online distillation (via collaborative learning) to transfer pose driven attention knowledge learned from VPN [30] to RGB stream. Distillation methods like [52], [53] are close to our approaches, however they are specific for image domain applications. In contrast, the extension of VPN: **VPN++** is dedicated for combining cross-modal information pertaining to video domain applications. The feature-level and attention-level distillation mechanisms to infuse Poses into RGB stream through cross-modal knowledge distillation provide a practical model for combining RGB and 3D Poses.

3 Video-Pose embedding Models

In this section, we first detail our previously proposed Video-Pose Network (VPN), followed by an elaborate description of the video-pose embedding models through distillation. We aim at building a video-pose network **VPN++** which benefits from two levels of distillation - (i) feature-level, and (ii) attention-level. At training time, the inputs to these models are RGB videos along with their corresponding 3D poses. These 3D poses could be obtained either from Kinect sensors using [54] or from RGB images using pose estimation algorithms like LCRNet++ [7] and VideoPose3D [8]. The RGB images and the 3D Poses are processed by a video backbone and a pose backbone respectively. In this work, the video backbones are usually 3D CNNs that take as input a stack of human cropped images from a video clip to compute the spatio-temporal representation of the clip. On the other hand, the Pose backbones are spatio-temporal Graph Convolutional Networks that take a stack of 3D Poses as a graphical input to model actions. At inference time, traditional VPN requires both RGB and 3D Poses to predict the actions. In contrast, VPN++ requires only the RGB videos at inference time to predict the action classes.

3.1 Background: VPN

VPN can be thought as a layer which can be placed on top of any 3D convolutional backbone. VPN takes as input a 3D feature map ($f \in \mathbb{R}^{c \times t \times m \times n}$) and its corresponding 3D poses (P) to perform two functionalities as shown in fig. 3. First, to provide an accurate alignment of the human joints with the feature map f. Second, to compute a modulated feature map (f') which is further classified for action recognition. The modulated feature map (f') is weighted along space and time as per its relevance. VPN exploits the highly informative 3D pose information to transform the visual feature map f and finally, compute the attention weights. This network has two major components: (I) an attention network and (II) a spatial embedding.

3.1.1 Attention Network

The attention network consists of a Pose Backbone and a spatio-temporal Coupler (STC). The input poses along the video are processed in a Pose Backbone as shown in fig 3. The pose based inputs of VPN are the 3D human joint coordinates $P \in \mathbb{R}^{3 \times J \times t_p}$ stacked along t_p temporal dimension, where J is the number of skeleton joints. The Pose Backbone processes these 3D poses to compute pose features h^* which are used further in the attention network for computing the spatio-temporal attention weights.

Next, the attention network in VPN learns the spatio-temporal attention weights A from the output of Pose Backbone in two steps as illustrated in fig. 4. In the first step, $m \times n$ dimensional spatial and t dimensional temporal attention weights are classically trained as in [40] to get the most important body parts and key frames for an action. This learning of spatial and temporal attention weights takes place in two streams (z_1 and z_2) consisting of dense layers each followed by relevant activations. In the second step, joint spatio-temporal attention weights are computed by performing a Hadamard product on the spatial and temporal attention weights. In order to perform this matrix multiplication, the spatial and temporal attention weights are inflated by duplicating the same attention weights in temporal and spatial dimension respectively.

This two-step attention learning process enables the attention network to compute spatio-temporal attention weights in which the spatial saliency varies with time. The obtained attention weights are crucial to disambiguate actions with similar appearance as they may have dissimilar motion over time. Finally, the spatio-temporal attention weights $A \in \mathbb{R}^{t \times m \times n}$ are linearly multiplied with the input video feature map f, followed by a residual connection with the original feature map f to output the modulated feature map f'. The residual connection enables the network to retain the properties of the original visual features.

3.1.2 Spatial Embedding of RGB and Pose

The objective of the embedding model is to provide tight correspondences between both pose and RGB modalities used in VPN. The state-of-the-art methods [15], [28] attempt to provide the attention weights on the RGB feature map using 3D pose information without projecting them into the same 3D referential. The mapping with the pose is only done by cropping the person within the input RGB images. The spatial attention computed through the 3D joint coordinates does not correspond to the part of the image (no pixel to pixel correspondence), although it is crucial for recognizing fine-grained actions. To correlate both modalities, an embedding technique inspired from image captioning task [55], [56] is used to build an accurate RGB-Pose embedding in order to enable the poses to represent the visual content of the actions.

Thus, the embedding is performed by propagating a normalized euclidean loss between the visual features and a spatial attention vector (z_1 obtained from STC). Both the visual feature and spatial attention vectors are obtained by

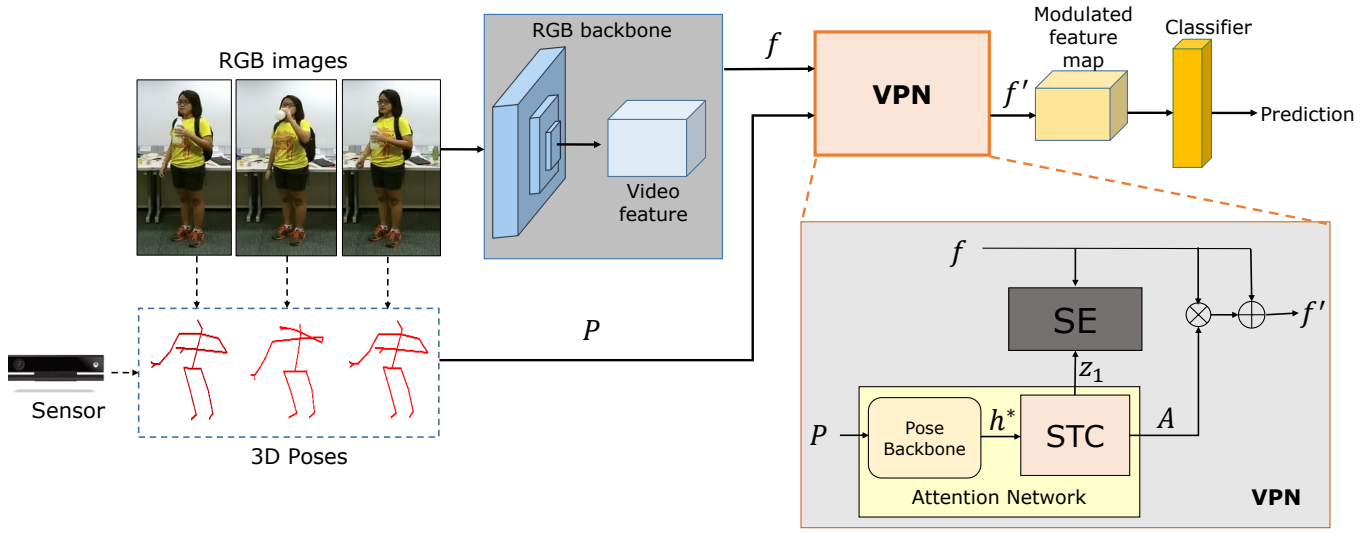

Fig. 3: VPN takes as input RGB images with their corresponding 3D poses. The RGB images are processed by a visual backbone which generates a spatio-temporal feature map (f). The proposed **VPN** takes as input the feature map (f) and the 3D poses (P). VPN consists of two components: an attention network and a spatial embedding (SE). The attention network further consists of a Pose Backbone and STC (spatio-temporal Coupler). VPN computes a modulated feature map f'. This modulated feature map f' is then used for classification.

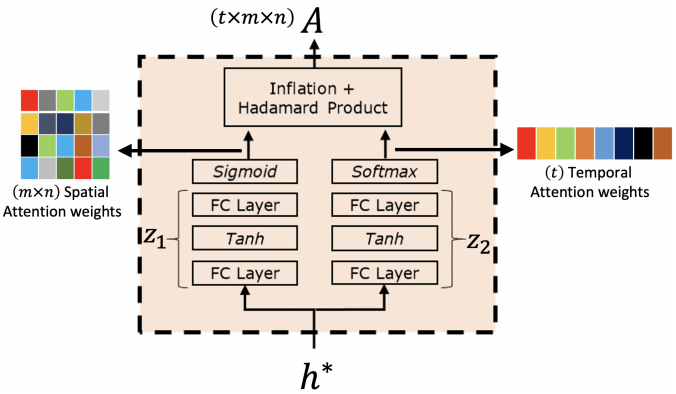

Fig. 4: STC: spatio-temporal Coupler to generate spatio-temporal attention weights A from the latent pose based feature h^*.

linear projection of the video content f and the 3D poses into a common dimensional embedding space.

Finally, VPN is plugged into the 3D ConvNet for an end-to-end training with a regularized loss L which is a convex combination of entropy loss, embedding loss and an attention regularization loss (refer to [30] for details).

Aiming at learning similar video-pose embeddings by hallucinating discriminative pose-level features, we propose an extension of VPN, namely VPN++. VPN++ effectively makes use of the pose features at training time and eliminates its reliance over Poses at inference time. In fig. 5, we provide a schematic diagram of VPN and our proposed distillation models to illustrate the disparities among them. VPN++ with only feature-level distillation is denoted as VPN-F and VPN++ with only attention-level distillation is denoted as VPN-A. Below, we elaborate the two levels of distillation in VPN++.

3.2 VPN-F (Feature-level distillation)

VPN++ involves knowledge distillation among modalities and thus, we have a teacher-student structure. This is referred to as the feature-level distillation in our model to infuse Pose stream into RGB stream. This is an attempt analogous to the spatial embedding in VPN. In order to perform this distillation, the Pose stream is considered as the Teacher Network $\mathcal{T}_F$, whereas the RGB stream as the Student Network $\mathcal{S}$. But unlike previous teacher-student networks [29], [49], here the teacher network occasionally provides irrelevant features, especially for actions where appearance information is important. For instance, using only Poses cannot discriminate actions like *wearing a shoe* or *taking off a shoe* but they can provide salient information about the localization of the action. For disambiguating these actions with similar appearance, we have to go beyond just mimicking the Pose stream to capture discriminative information. Consequently, we use the concept of contrastive learning to learn a representation for which the positive pairs are close to each other and negative pairs are pushed apart in some metric space. Most related to our work, Contrastive Representation Distillation [52] (CRD) involves learning an unsupervised representation through knowledge distillation followed by a downstream training on the same set of training samples. In contrast, we focus specifically on video domain (with RGB and 3D poses) and formulate a supervised training strategy. This strategy includes jointly optimizing the student network with the class labels $\hat{Y}$ in addition to distilling the knowledge from Pose stream to RGB. This enables the actions with similar appearance to move apart in the feature space due to their dissimilar distillation through pose embeddings. We call the model with only this feature-level distillation as VPN-F.

At training time, we learn the VPN-F representation in two steps. Let V_i be a video (stack of RGB frames) and P_i be the

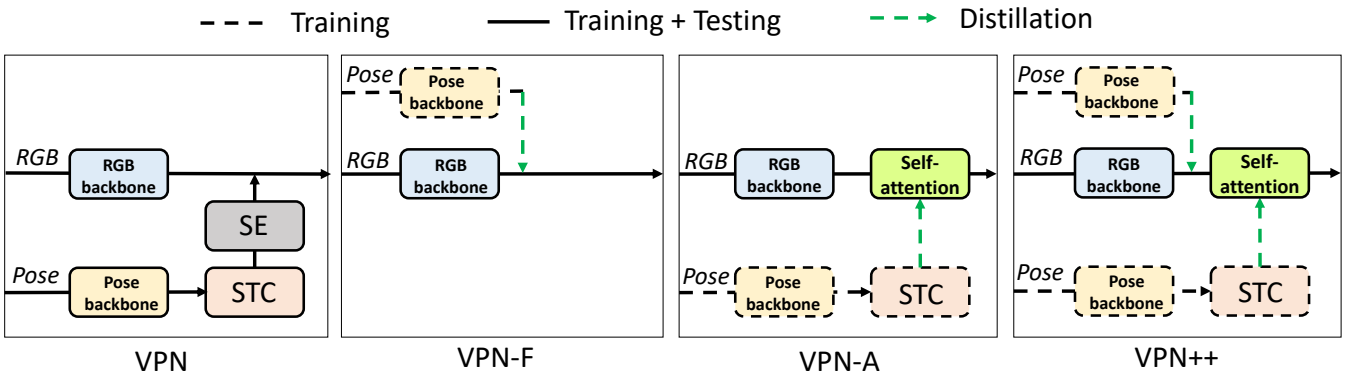

Fig. 5: A schematic diagram of our models - VPN, VPN-F, VPN-A, and VPN++ reflecting their variation for providing video-pose embeddings. The inputs to each model at training are the *RGB* and *Poses*. STC indicates the spatio-temporal coupler learning the attention weights. SE indicates the spatial embedding module.

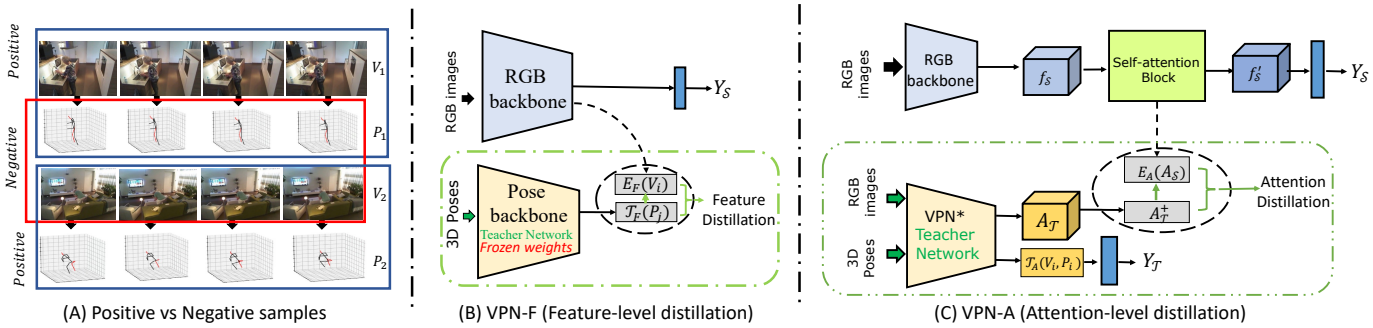

Fig. 6: (A) The positive & negative video-pose pairs (at the left) are input to the teacher-student network. (B) **VPN-F**: VPN++ distillation model with only feature-level distillation. Here, the Pose Teacher network is pre-trained for action classification. Supervised Contrastive Distillation (SCD) is applied between the RGB and Pose features. (C) **VPN-A**: VPN++ distillation model with only attention-level distillation. Here, the teacher VPN* is the video-pose network [30] without the spatial embedding (SE). Also, $A_{\mathcal{T}}$ and $\mathcal{T}_A(V_i, P_i)$ can be referred to as the attention weights (A) and modulated feature map (f') of VPN (see fig. 3). Teacher network VPN* is trained collaboratively with the student RGB backbone.

corresponding 3D Poses for the i^{th} sample in the training set. In the first step, the teacher network $\mathcal{T}$ is trained with the 3D poses for classifying V_i into $\mathcal{C}$ action classes and its weights are then frozen.

In the second step, our goal is to learn a latent space where semantically related RGB frames and Poses are close to each other and far away otherwise. We achieve this by imposing a supervised contrastive distillation (SCD) loss between the teacher and the student at the feature level as illustrated in fig. 6 (B). Inspired from audio-video [57] and text-video analysis [58], we assign a set of candidate positive pairs (V_i, P_i), thus the RGB frames and 3D Poses are extracted from the same video labeled action $C_k \in \mathcal{C}$. On the other hand, the negative pairs are some randomly associated data (V_i, P_j) where P_j is randomly chosen from the subset $\mathcal{C} \setminus C_k$ as shown in fig. 6 (A).

For distillation, the SCD loss is imposed between the features at the output of the layer immediately before the final fully-connected layer of the teacher network and the features of the visual embedding obtained from the RGB student network. This visual embedding $E_F(V_i)$ is a linear projection of f_S, where spatio-temporal feature map f_S is computed by the RGB backbone $\mathcal{S}(V_i)$. We denote the features from the teacher network as $\mathcal{T}_F(P_j)$. We maximize the mutual information between Pose teacher and RGB student representations by jointly optimizing the student network at the same time as we learn a video-pose embedding $[\mathcal{T}_F(P_j), E_F(V_i)]$. Thus, our distillation loss over a batch of data ($\mathcal{B}$) is formulated as the log likelihood of the data under this model:

$$\mathcal{L}_{SCD} = \frac{1}{|\mathcal{B} - \mathcal{N}|} \sum_i \log[\mathcal{T}_F(P_i), E_F(V_i)] + \sum_{j \neq i} \log(1 - [\mathcal{T}_F(P_j), E_F(V_i)]) \quad (1)$$

$$\text{where} \quad [\mathcal{T}_F(P_j), E_F(V_i)] = \frac{e^{\mathcal{T}_F(P_j)^\intercal E_F(V_i)}}{e^{\mathcal{T}_F(P_j)^\intercal E_F(V_i)} + \mathcal{M}}$$

Here, $[\mathcal{T}_F(P_j), E_F(V_i)] \rightarrow (0, 1)$ corresponds to the video-pose embedding and constant $\mathcal{M}$ is determined by the ratio of the number of negatives N to the cardinality of the dataset. Thus for the positive pairs, $\mathcal{L}_{SCD}$ enforces the video student representation $E_F(V_i)$ to project along the Pose teacher representation $\mathcal{T}_F(P_i)$. Conversely for the negative pairs, the student representation is projected perpendicular to the teacher representation in feature space. This feature modulation (at student network) due to the distillation loss

is accompanied by cross-entropy loss $\mathcal{L}_C^{\mathcal{S}}$ to optimize the RGB student network for predicting the action labels Y_S. This joint optimization induces a selective infusion of the pose features in the RGB space with respect to the action class. Note that the student network is always fed with the ground-truth $\hat{Y}$ corresponding to the RGB input. So, a video sample with corresponding ground-truth is repeated twice in a mini-batch while training VPN-F with SCD loss.

3.3 VPN-A (Attention-level distillation)

Next, we aim at learning RGB representation benefiting from attention mechanism. Attention mechanisms focusing on salient image region across time have become instrumental for discriminative visual representation. For RGB based ADL recognition, VPN [30] has shown that pose driven attention mechanism is more accurate and effective compared to the ones using self-attention mechanisms through RGB itself. Therefore, we develop a second-level of distillation for transferring pose driven attention knowledge to RGB stream. For the sake of simplicity, we first explain the model with only attention-level distillation, dubbed as VPN-A. For this distillation, we chose our Video-Pose Network [30] as a Teacher network $\mathcal{T}_A$. This Video-Pose Network (VPN*) is implemented following [30] with no spatial embedding since we find that the feature-level distillation could hallucinate the features learned through spatial embedding in VPN. The student network $\mathcal{S}$ is a video backbone, similar to the one used in VPN-F.

The challenge is to transfer the knowledge of attention weights learned by the teacher to the RGB student network. Therefore, a self-attention block similar to [9] is invoked in the RGB based network which could learn the attention weights from the teacher. However, a feature-level distillation in this case does not activate the relevant neurons at the student network. We empirically support this claim in the experimental analysis. Moreover, learning attention weights is an evolutionary mechanism where a model learns the salient regions in the spatio-temporal space with every batch of iteration over the training data. So, for this level of distillation, we opt for online distillation, where the teacher VPN* and the student RGB backbone along with the self-attention block collaboratively optimize their respective entropy loss as illustrated in fig. 6 (C). Such a distillation encourages the RGB student to produce similar attention weights as the VPN* teacher, intuitively paying attention to similar parts of the video as the teacher.

VPN-A is trained in a single step. On one hand, VPN* teacher network is trained with action labels to learn pose driven attention weights $A_{\mathcal{T}}$ to modulate its RGB feature map. Note that these attention weights corresponds to A from STC of VPN discussed in section 3.1. On the other hand, the student network intakes only the RGB frames. The self-attention block projects the RGB feature $f_{\mathcal{S}}$ to a query (Q) and memory (key and value, K & V) embedding using linear projections ($1 \times 1 \times 1$ Conv), where typically the query and keys are of lower dimension (see fig. 7 for a zoom into the self-attention block). The output for the query, i.e. the modulated feature map $f'_{\mathcal{S}}$, is computed as an attention weighted sum of values V, with the attention weights $A_{\mathcal{S}}$ obtained from the product of the query Q with keys K. The

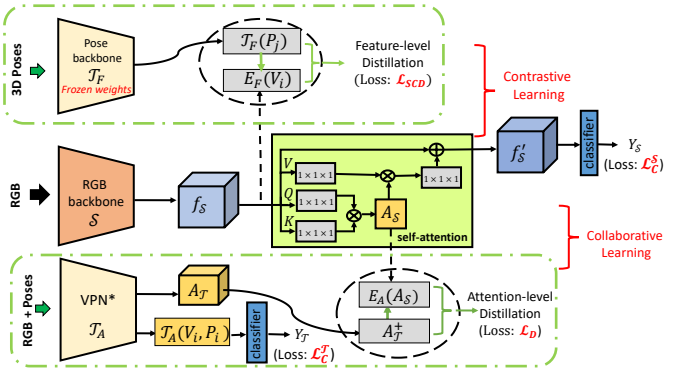

Fig. 7: **VPN++**: The proposed distillation model when both VPN-F and VPN-A are integrated into a single model. The student network consists of a RGB backbone and a self-attention bock. At training, the model is trained in a contrastive manner for the feature-level distillation, and collaborative manner for the attention-level distillation. Note that Video-Pose attention model VPN* does not have the spatial embedding module.

attention weights $A_{\mathcal{S}}$ have to be learned from the evolution of 3D Poses. So, we invoke a distillation loss between the self-attention and VPN attention weights ($A_{\mathcal{S}}$ & $A_{\mathcal{T}}$). The distillation loss is a classical Mean Squared Error (MSE) loss between the attention weight embeddings ($E_A(A_{\mathcal{S}})$ & $A_{\mathcal{T}}^+$) from the teacher-student network. The projection of the attention weights ($A_{\mathcal{S}}$ & $A_{\mathcal{T}}$) is necessary since they differ in terms of their dimensionality. This projection is performed by linearly transforming them into the same dimension and further normalizing them through $\mathcal{L}$-2 norm. The distillation loss $\mathcal{L}_{\mathcal{D}}$ is formulated as

$$\mathcal{L}_D = ||A_{\mathcal{T}}^+ - E_A(A_{\mathcal{S}})||^2 \tag{2}$$

The VPN* backbone, i.e. $\mathcal{T}_A(V_i, P_i)$ classifies its modulated feature map using the entropy loss $L_C^{\mathcal{T}}$ between the true class labels $\hat{Y}$ and the predicted class labels $Y_{\mathcal{T}}$. Besides, the modulated feature map $f'_{\mathcal{S}}$ at the student network is classified simultaneously using the entropy loss $L_C^{\mathcal{S}}$ between the same true class labels and the predicted class labels $Y_{\mathcal{S}}$.

3.4 VPN++: Integrating VPN-F & VPN-A

Finally, we aim at learning a unified RGB representation that can emulate both Pose based features and pose driven attention weights. This objective also encourages the model to jointly optimize the two levels of distillation loss along with the cross entropy loss to learn the class labels. Thus, we integrate the two levels of distillation into a single model - we call VPN++. The training methodology of VPN++ involves contrastive learning for the feature-level distillation and collaborative learning for the attention-level distillation. In fig. 7, we show the VPN++ model with two levels of distillation. Here the RGB student network includes the RGB backbone and the self-attention block whereas there are two teacher networks - a pre-trained Pose backbone for infusing the pose features to the RGB stream, and VPN* for transferring the pose driven attention knowledge to the self-attention block of the student network. In order

to incorporate the contrastive and collaborative learning strategies both in the same model, a batch of samples with (positive, negative) pairs for the feature-level distillation and (positive, positive) pairs for the attention-level distillation is fed to the model. Note that the Pose teacher network for feature-level distillation is frozen.
Thus, the RGB student network is jointly optimized with the following linear combination of the distillation losses and the entropy losses:

$$\mathcal{L} = \mathcal{L}_C^{\mathcal{S}}(Y_S, \hat{Y}) + \mathcal{L}_C^{\mathcal{T}}(Y_{\mathcal{T}}, \hat{Y}) - \alpha\mathcal{L}_{SCD} + \beta\mathcal{L}_D \quad (3)$$

where α and β are the weighting factors of the distillation losses. Thus, VPN++ not only learns to distill the pose knowledge into RGB but also learn discriminative representation through pose driven attention distillation. While testing VPN++ (the RGB student network), we only use RGB frames as input to compute the action class scores, avoiding the requirement of 3D Poses.

4 Experiments

We evaluate the effectiveness of VPN++ and its corresponding components for action classification on four datasets popular for ADL: a real-world dataset - Toyota-Smarthome [15], a large scale human activity dataset - NTU RGB+D-60 [16], the super-set of NTU-60 dataset - NTU RGB+D-120 [17], and a relatively small scale human-object interaction dataset - Northwestern-UCLA [59].
Toyota-Smarthome (Smarthome or SH) is a recent ADL dataset recorded in an apartment where 18 older subjects carry out tasks of daily living during a day. The dataset contains 16.1k video clips, 7 different camera views and 31 complex activities performed in a natural way without strong prior instructions. This dataset provides RGB data and 3D skeletons which are extracted from LCRNet [7]. For evaluation on this dataset, we follow cross-subject (CS) and cross-view (CV_2) protocols proposed in [15]. We ignore protocol CV_1 due to limited training samples.
NTU RGB+D (NTU-60 & NTU-120): NTU-60 is acquired with a Kinect v2 camera and consists of 56880 video samples with 60 activity classes. The activities were performed by 40 subjects and recorded from 80 viewpoints. For each frame, the dataset provides RGB, depth and a 25-joint skeleton of each subject in the frame. For evaluation, we follow the two protocols proposed in [16]: cross-subject (CS) and cross-view (CV). NTU-120 is a super-set of NTU-60 adding a lot of new similar actions. NTU-120 dataset contains 114k video clips of 106 distinct subjects performing 120 actions in a laboratory environment with 155 camera views. For evaluation, we follow a cross-subject (CS_1) protocol and a cross-setting (CS_2) protocol proposed in [17].
Northwestern-UCLA Multiview activity 3D Dataset (N-UCLA) is acquired simultaneously by three Kinect v1 cameras. The dataset consists of 1194 video samples with 10 activity classes. The activities were performed by 10 subjects, and recorded from three viewpoints. We performed experiments on N-UCLA using the cross-view (CV) protocol proposed in [59]: we trained our model on samples from two camera views and tested on the samples from the remaining view. For instance, the notation $V^3_{1,2}$ indicates that we trained on samples from view 1 and 2, and tested on samples from view 3.

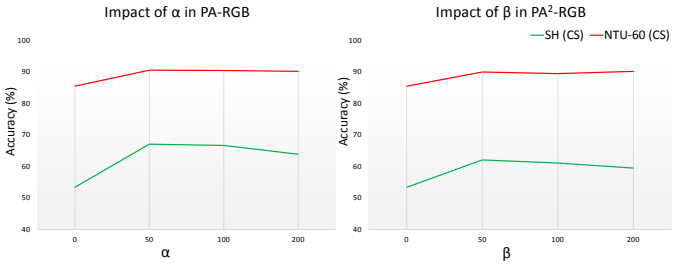

Fig. 8: Accuracy of VPN-F (on left) and VPN-A (on right) for different values of α & β respectively on Smarthome (CS) and NTU-60 (CS) datasets.

4.1 Implementation details

For the input at training time, the 3D Poses are provided for NTU and N-UCLA dataset. For Smarthome dataset, two sets of 3D Poses, namely old and new Poses, are provided which are eventually extracted from RGB. Note that the new 3D Poses are of higher quality compared to the older ones.
For VPN++, the Teacher network for feature-level distillation is AGCN-J [20] Pose backbone. Thus, we follow the pre-processing step on the 3D Poses as in [20]. For attention-level distillation, the Teacher Network (VPN*) is adapted with a 2 layer AGCN as Pose backbone and no spatial embedding. The Student network is I3D [3] RGB backbone pre-trained on ImageNet [60] and Kinetics-400 [5]. It takes 64 RGB frames as input. The self-attention block is implemented with an additional Non-Local block [9] placed on top of the I3D (`Mixed_5c` layer).
Training. For training the teacher networks of VPN++ with categorical cross entropy loss, we follow the steps as in [20] and [30]. For training the student network, a dropout [61] of 0.3 and a *softmax* layer are added at the end of the self-attention block for class prediction. VPN++ is trained with a 4-GPU machine where each GPU has 4 video clips in a mini-batch. It is trained with SGD optimizer having initial learning rate of 0.01, momentum of 0.9, and a weight decay rate of 0.1 after every 10 epochs. While training VPN++, we chose $\alpha = \beta = 50$. For feature-level distillation, each batch consists of 8 positives and 8 negatives.
Inference. At test time, we perform fully convolutional inference in space as in [9]. The final classification is obtained by max-pooling the softmax scores.

4.2 Hyper-parameter sensitivity

VPN-F and VPN-A distillation models are trained by a linear combination of two losses: cross-entropy loss between the logits and the ground-truth targets, and the distillation loss between the video-pose features. In fig. 8, we report the accuracy of VPN-F and VPN-A on Smarthome and NTU-60 datasets using different values of α and β respectively. We observe that a non-zero value of α or β increases the action classification accuracy compared to the baseline RGB stream. This shows the importance of the distillation from Pose stream to RGB in both models. We also observe that by increasing the weighting factor of the distillation loss, we reach a peak accuracy for $\alpha = \beta = 50$ as shown in fig. 8. This

Loss	SH (CS)	NTU-60 (CS)	NTU-60 (CV)
MSE [50]	61.8	89.1	92.4
CRD [52]	64.7	90.1	93.1
SCD	**67.1**	**90.8**	**93.8**

TABLE 1: Ablation for choice of distillation loss in VPN-F.

Model	Poses	SH (CS)	NTU-60 (CS)	NTU-60 (CV)
I3D w/o attention (backbone)	×	53.4	85.5	87.3
I3D w NL attention (self-attn)	×	53.6	88.4	87.1
I3D w pose attention (VPN)	✓	**65.2**	**93.5**	**96.2**

TABLE 2: Impact of pose driven attention (VPN) compared to RGB based Non Local (NL) attention mechanism.

shows that our distillation models effectively leverage both RGB and Pose streams to classify the action when combined in a strategic manner. Further increase in the values of α or β influences the distillation loss to dominate the student RGB network while training. This causes the resultant student network to mimic the Pose stream rather than exploiting both streams.

For SCD loss, the choice of number of negatives for each positive input (video-pose pair) is flexible. Conventionally, more negatives for each Positive in contrastive learning yields higher accuracy. This observation is not noted in our case due to a supervised strategy of using the contrastive loss. Thus, we take one negative for each positive to train VPN-F. We utilize the above observation for hyper-parameters while training VPN++.

4.3 Ablation studies

In this section, we analyze the impact of proposed distillation methods w.r.t. previous methods. We also quantify the robustness of VPN-F and VPN-A.

Which loss is better for feature-level distillation? In this ablation study (Table 1), we compare different distillation losses for transferring knowledge from pose features to RGB features. The training strategy for all these losses are different but are applied between the video-pose features $\mathcal{T}_F(P_j)$ and $E_F(V_i)$. The mechanism of learning visual representation with the concept of contrastive learning between the positive and negative samples (CRD [52] and SCD) outperform the classical way to distillate knowledge using MSE loss [50]. We also note that our SCD outperforms CRD significantly on Smarthome, whereas the margin of improvement on NTU is comparatively low. This indicates that CRD is effective for scenarios where high quality Poses are available and SCD is consistently effective even for low quality Poses.

Why do we need to emulate pose driven attention? We know that attention mechanisms are crucial for understanding ADL [15]. But attention weights obtained using RGB based self-attention mechanism like Non-Local blocks [9] rely too much on variation of intensities in spatio-temporal feature maps, hence lacks semantics. In contrast, 3D poses capture the semantics in the videos and significantly improve the action recognition performance as shown in Table 2. Our VPN with pose driven attention significantly improves the action classification accuracy by relatively 20.8%

Loss & distillation strategies	Collaborative learning	SH (CS)	NTU-60 (CS)	NTU-60 (CV)
SCD (feature level)	×	55.7	89.1	91.5
SCD (feature level)	✓	51.1	87.1	89.5
SCD (attention weights)	×	54.2	88.9	91.4
SCD (attention weights)	✓	53.1	87.4	90.6
MSE (attention weights)	×	61.1	89.1	92.4
MSE (attention weights)	✓	**62.1**	**90.0**	**93.1**

TABLE 3: Comparison of VPN-A with other strategies to distill pose driven attention.

Combining strategy	L_e	Test time (s)	SH CS	NTU-60 CS	NTU-60 CV
VPN-F (1^{st}) + VPN-A (2^{nd})	×	0.4	62.3	90.5	93.2
VPN-A (1^{st}) + VPN-F (2^{nd})	×	0.4	63.9	90.6	93.4
VPN-F + VPN-A (Late Fusion)	×	0.7	68.7	91.7	94.8
VPN++ (multi-teacher)	✓	0.4	68.9	**91.9**	94.8
VPN++ (multi-teacher)	×	0.4	**69.0**	**91.9**	**94.9**

TABLE 4: Comparison of different strategies to combine VPN-F & VPN-A. L_e represents the spatial embedding in VPN* (teacher of VPN-A).

on Smarthome dataset. It is worth noting that the improvement is significant for Smarthome compared to NTU-60 as it contains many fine-grained actions with videos captured by fixed cameras in an unconstrained Field of View. Thus, enforcing the embedding loss enhances the spatial precision during inference. Therefore, we chose to mimic pose driven attention for a second-level of distillation in VPN++.

Which loss is better for attention-level distillation? For VPN-A, we propose to distill knowledge at attention-level than at feature-level due to its more effectiveness as supported by the experiments in Table 3. Note that the feature-level distillation in the former experiment is performed between the output of the modulated feature map of VPN*, i.e. $\mathcal{T}_A(V_i, P_i)$ and the modulated feature map of the student network f'_S. We investigate the effectiveness of collaborative training the teacher-student network for transferring attention-level features. In these experiments in Table 3, the VPN* teacher network is pre-trained and frozen when collaborative training is not performed. We also compare the performance of supervised contrastive distillation loss (most effective loss in Table 1) with MSE loss at attention-level. However, MSE loss with collaborative training strategy to distill attention weights from VPN* Teacher network to RGB based Non-Local student network outperforms the baselines by up to 7.9% on Smarthome dataset. This shows that reducing MSE between the attention weights of video and pose embeddings is a better strategy to distill attention weights than contrastive learning. This is coherent with the fact that distillation of attention weights do not correspond to positives and negatives w.r.t. video samples whereas distillation at feature level represents entities that could have positives and negatives.

How to combine VPN-F & VPN-A? In Table 4, we observe a performance drop when both the training strategies of VPN-F & VPN-A are combined in a sequential manner, one after the other. The cause for this performance drop is the difficulty for the second distillation to significantly modify the RGB feature map (at student's network) once the first distillation has modified it. In contrast to these fusion strate-

	Stream	SH (CS)	NTU-60 (CS)	NTU-60 (CV)	NTU-120 (CS_1)	NTU-120 (CS_2)	N-UCLA ($V^3_{1,2}$)
	l_1: RGB	53.4	85.5	87.3	77.0	80.1	86.0
	l_2: 3D Poses	51.5	85.8	93.8	79.6	81.1	78.2
	$l_1 + l_2$ (Late Fusion)	63.0	87.7	94.8	81.1	83.3	87.1
	$l_1 + l_2$ (attention)	65.2	93.5	96.2	86.3	87.8	93.5
Ours	VPN-F	67.1	90.8	93.8	85.1	87.6	89.1
Ours	VPN-A	62.1	90.0	93.1	85.2	88.0	88.2
Ours	VPN++	69.0	91.9	94.9	86.7	89.3	91.9
Ours	VPN++ + 3D Poses	**71.0**	**94.9**	**98.1**	**90.7**	**92.5**	**93.5**

TABLE 5: Top-1 accuracy of RGB, 3D Poses, VPN-F, VPN-A, and VPN++ on 4 datasets.

Dataset	Pose Quality	AGCN-J [20]	VPN [30]	VPN++
SH (CS)	Medium	54.0	65.2	**69.0**
SH (CS)	Low	49.1	62.1	**66.8**
NTU-60 (CS)	High	85.8	**93.5**	91.9
NTU-60 (CS)	Low	44.4	90.1	**91.3**

TABLE 6: Performance of several methods with different levels of pose quality.

gies, the score level fusion of both student networks significantly outperforms the above two end-to-end strategies. Similar improvement is noted for our multi-teacher network trained with contrastive and collaborative strategy. Due to the lower final model complexity and lower inference time, the multi-teacher network is superior than the one with late fusion. This performance improvement highlights the complementary optimizations which are well preserved while training jointly in a single model in VPN++. Finally, we also observe that spatial embedding in the teacher network of attention-level distillation do not contribute to the classification accuracy and hence can be ignored. This shows that the feature-level distillation could hallucinate the pose features performed by the spatial embedding in VPN.

Comparison of distilled models with RGB & Pose streams. In Table 5, we compare our distillation models - VPN-F & VPN-A with uni-modal models and their combinations. RGB and 3D Poses are modeled using I3D [3] and AGCN-J [20] networks. Following the state-of-the-art trends, RGB and Poses are combined using score level fusion (Late Fusion) and attention mechanism (VPN). Both VPN-F & VPN-A significantly outperform the individual modalities. VPN-F with contrastive distillation outperforms the late fusion strategy of combining RGB and Poses on all the datasets except NTU-60 (CV protocol). This exception is coming from the high action classification performance (93.8%) with view-invariant 3D poses for cross-view protocol of NTU-60, where high quality 3D Poses are available. On the other hand, VPN-A is an attention based model and requires subsequently large amount of data for learning salient attention weights. This is corroborated by its lower classification accuracy for NTU-60 in contrast to NTU-120 (up to 0.4% higher than even VPN-F) where the number of training samples is two times that of NTU-60. The combination of VPN-F & VPN-A in VPN++ further boosts the classification accuracy by up to 2.8% relatively on Smarthome. Further improvement in action classification when combined with 3D Poses indicates that our distilled models still lacks in terms of mimicking the Pose stream. However, their superior performances compared to all prior techniques of combining RGB and 3D Poses show the discriminative representation learned by our VPN++.

What happens when the Pose quality degrades? As shown in fig. 2, VPN++ does not require Poses at test time which substantially reduces the model inference time. Thus, the bad quality of Poses at inference time does not hamper the performance of these models. But what if the Poses are shoddy for the entire data distribution (even at training)? In Table 6, we dig deeper into this problem by investigating the influence of Pose quality on the performance for different models. First, we describe the experimental setup to obtain the different levels of Pose quality.

NTU-60 dataset was recorded in a laboratory, so we can have **hight-quality 3D Poses** captured by the Microsoft Kinect v2 sensor. For **low-quality 3D Poses**, we downscaled the original videos by reducing their resolution to 320×180 and randomly invoke partial occlusions to fabricate the dataset similar to real-world settings. Then, we extract the 3D Poses using LCRNet++ [7]. In Smarthome, Poses are obtained from RGB rather than using depth-map. For **medium-quality 3D Poses**, we apply Selective Spatio-Temporal Aggregation based Pose Refinement System (SSTA-PRS) [62] which aims at improving the performance of pose estimation by integrating the advantages of several state-of-the-art pose estimation systems (eg. LCRNet++ [7], OpenPose [63] and AlphaPose [64]) to extract 2D Poses. Then, we apply VideoPose3D [8] to obtain 3D Poses over 2D Poses. For the **low-quality 3D Poses**, we only use LCRNet++ [7].

The two baselines compared with VPN++ in Table 6 include skeleton based model: AGCN-J [20] and RGB+Pose based attention model: VPN. We observe that VPN++ is less sensitive to the quality of Poses with a deterioration of classification accuracy by 3.1% and 0.6% on Smarthome and NTU-60 respectively compared to the baselines (4.7% & 3.6% for VPN and 9% & 48.2% for AGCN-J). This experiment shows that the quality of Poses highly impacts skeleton based action recognition, whereas our distillation model VPN++ outperforms even VPN without the requirement of Poses at inference. The tolerance of VPN++ to noisy Poses is due to the selective distillation of Pose features within the video-pose embedding of the distillation mechanisms. For instance, the degraded Poses contingent upon occlusions, low subject resolution, or other real world scenarios provide ambiguous features pertaining to actions. Thanks to the distillation mechanisms, we can filter the appropriate pose based features while infusing knowledge into the RGB stream.

5 Qualitative Visualization

In fig. 9, we present a visualization of class activation maps of RGB, VPN-F, VPN-A, and VPN++ using Grad-CAM [65]. These maps enable us to visualize discriminative regions specific to each action class. VPN-F, for actions like *reach into pocket*, VPN-A for actions like *clean dishes*, and both VPN-F & VPN-A for actions with subtle motion like *stirring*, focus sharply around the hands grasping objects providing contextual information worth modeling the actions. The activation map of RGB stream either focuses only on irrelevant

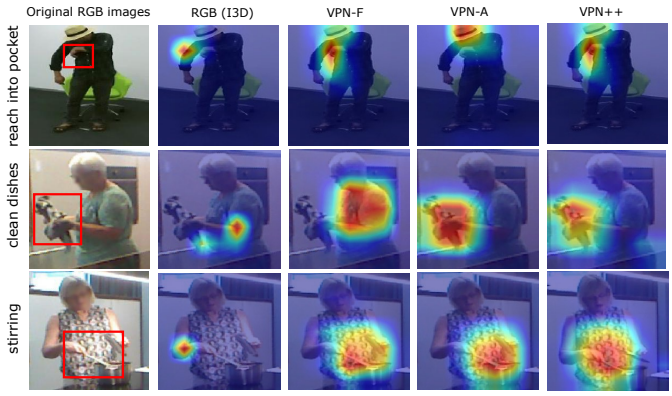

Fig. 9: Qualitative visualization of class activation maps of RGB, VPN-F, VPN-A, and VPN++ using Grad-CAM [65]. The red bounding box refers to the precised Region of Interest relevant to classifying the action.

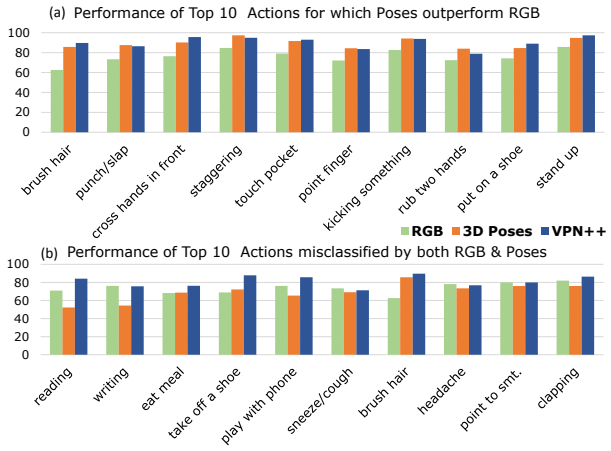

Fig. 10: Action classification accuracy (in %) of top-10 actions (a) for which Pose stream outperforms RGB stream, and (b) which are mis-classified by both RGB and Poses.

motion patterns (see fig. 9). Moreover, the class activation maps obtained for VPN++ select the ones (VPN-F or VPN-A) that is effective for an action class. Thus, our combining strategy of the two levels of distillation in VPN++ can take advantage of the complementary features learned via both distillation mechanisms. This qualitative visualization shows that our distillation mechanisms learn discriminative representation that exploits the contextual information in the scene which is crucial for ADL.

6 Quantitative Analysis

In this section, we present an analysis of Top-10 class-wise performance of an RGB based approach, a Pose based approach, & our distillation model VPN++ (see fig. 10). First, (a) we present the performance of Top 10 actions for which Poses outperform RGB. VPN++ outperforms the Pose stream for all these actions which shows that our model not only learns Pose based features but also learns an augmented representation from the Pose teacher network. Second, (b) we present the performance of Top

	Methods	Pose	RGB	Att	CS	CV_2
	DT [66]	×	✓	×	41.9	23.7
	I3D [3]	×	✓	×	53.4	45.1
	I3D+NL [9]	×	✓	✓	53.6	43.9
	AssembleNet++ [44]	×	✓	✓	63.6	-
Old Poses	LSTM [67]	✓	×	×	42.5	17.2
Old Poses	P-I3D [28]	✓	✓	✓	54.2	50.3
Old Poses	Separable STA [15]	✓	✓	✓	54.2	50.3
Old Poses	VPN [30]	✓	✓	✓	60.8	53.5
Old Poses	**VPN++**	○	✓	✓	**66.8**	**53.6**
New Poses	2s-AGCN [20]	✓	×	×	57.1	49.7
New Poses	VPN [30]	✓	✓	✓	65.2	54.1
New Poses	**VPN++**	○	✓	✓	69.0	54.9
New Poses	**VPN++ + 3D Poses**	✓	✓	✓	**71.0**	**58.1**

TABLE 7: Results on Smarthome dataset with cross-subject (CS) and cross-view (CV_2) settings (accuracies in %); Att indicates attention mechanism, ○ indicates that the modality has been used only in training.

	Methods	Pose	RGB	Att	CS	CV
	2s-AGCN [20]	✓	×	×	88.5	95.1
	DGNN [19]	✓	×	×	89.9	96.1
	MS-G3D Net [68]	✓	×	×	91.5	96.2
	PEM [69]	✓	✓	×	91.7	95.2
I3D	Separable STA [15]	✓	✓	✓	92.2	94.6
I3D	P-I3D [28]	✓	✓	✓	93.0	95.4
I3D	VPN [30]	✓	✓	✓	93.5	96.2
I3D	**VPN++**	○	✓	✓	91.9	94.9
I3D	**VPN++ + 3D Poses**	✓	✓	✓	**94.9**	**98.1**
RNX3D	VPN (RNX3D101) [30]	✓	✓	✓	95.5	98.0
RNX3D	RNX3D101+MS-AAGCN [21]	✓	✓	×	96.1	99.0
RNX3D	**VPN++**	○	✓	✓	93.5	96.1
RNX3D	**VPN++ + 3D Poses**	✓	✓	✓	**96.6**	**99.1**

TABLE 8: Results on NTU-60 dataset with cross-subject (CS) and cross-view (CV) settings (accuracies in %).

10 actions mis-classified by both RGB and Pose streams. Interestingly, VPN++ improves the performance of RGB stream for actions which are mostly mis-classified owing to two challenges - (i) similarity in appearance like *taking off a shoe* (+5%) or *wearing a shoe*, *clapping* (+4%) or *rubbing two hands*, and (ii) subtle motion while performing the actions like *reading* (+13%), *writing* (+1%), and *headache* (+2%). Thus, VPN++ confirms empirically its potential to mitigate the drawbacks of SoA approaches by effectively providing an appropriate combination of the modalities (RGB and Poses) through distillation.

7 Comparison to the the State-of-the-Art

We compare VPN++ to the State-of-the-Art (SoA) on Smarthome, NTU-60, NTU-120, and N-UCLA in Tables 7, 8, 9, and 10.

For smarthome dataset, we present the SoA categorized into RGB and RGB+Pose based methods in Table 7. We provide the evaluation results on old Poses and new Poses (referred to as low & medium levels of Pose quality in Table 6). VPN++ outperforms all the SoA methods by up to 9.8% & 5.8% relatively with old and new Poses respectively. This significant improvement on this dataset can be

Methods	Pose	RGB	Att	CS_1	CS_2
2s-Att LSTM [70]	✓	×	✓	61.2	63.3
Multi-Task CNN [71]	✓	×	×	62.2	61.8
PEM [69]	✓	×	✓	64.6	66.9
2s-AGCN [20]	✓	×	✓	82.9	84.9
MS-G3D Net [68]	✓	×	×	86.9	88.4
Two-streams [1]	×	✓	×	58.5	54.8
I3D* [3]	×	✓	×	77.0	80.1
Two-streams + ST-LSTM [17]	✓	✓	×	61.2	63.1
Separable STA* [15]	✓	✓	✓	83.8	82.5
VPN [30]	✓	✓	✓	86.3	87.8
VPN++	○	✓	✓	86.7	89.3
VPN++ + 3D Poses	✓	✓	✓	**90.7**	**92.5**

TABLE 9: Results on NTU-120 dataset with cross-subject (CS_1) and cross-setup (CS_2) settings (accuracies in %); Att indicates attention mechanism.

Methods	Data	Att	$V^3_{1,2}$
HPM+TM [72]	Depth	×	91.9
Ensemble TS-LSTM [73]	Pose	×	89.2
SGN [74]	Pose	×	92.5
NKTM [75]	RGB	×	85.6
I3D* [3]	RGB	×	86.0
Glimpse Cloud [14]	RGB+ $\overline{Pose}$	✓	90.1
Separable STA [15]	RGB+Pose	✓	92.4
P-I3D [28]	RGB+Pose	✓	93.1
Global Model [76]	RGB+Pose	✓	**93.5**
VPN [30]	RGB+Pose	✓	**93.5**
VPN++	RGB+ $\overline{Pose}$	✓	91.9
VPN++ + 3D Poses	RGB+Pose	✓	**93.5**

TABLE 10: Results on N-UCLA dataset with cross-view $V^3_{1,2}$ settings (accuracies in %); $\overline{Pose}$ indicate its usage only in the training phase.

explained by the video-pose embedding infused through our two levels of distillation for combining RGB & Poses, which in turn handles the challenge of low camera framing [15]. As discussed earlier, often low quality Poses are obtained in real-world scenarios with occlusions and low subject resolution. Thanks to the distillation mechanisms, it encourages the classification model to selectively infuse the relevant Pose information into the RGB stream. by providing a discriminative video-pose embedding. For NTU-60 dataset, VPN++ achieves accuracy close to the methods requiring Poses at test time whereas for NTU-120, VPN++ outperforms the later. We observe that the skeleton based action recognition methods perform better compared to the RGB based methods on NTU dataset. But this is due to the high quality of Poses (with no occlusion) which makes the dataset apt for Pose only methods. On the contrary, in real-world dataset like Smarthome (see Table 7), the Pose only methods substantially under-perform compared to the RGB based methods. Another limitation of Pose only methods includes their lack of appearance encoding. However, VPN++ when combined with 3D Poses outperforms SoA on both NTU datasets. We confirm the robustness of VPN++ by evaluating it with 3D ResNext-101 [34] as a video backbone on NTU-60. Similar observations can also be done on N-UCLA dataset in Table 10 hinting that VPN++ generalizes over small scale datasets too.

Settings Criterion	Model Choice	Dataset	Inference time	# Param.	Acc. (in %)
Baseline (RGB)	I3D [3]	SH	0.3s	12M	53.4
Baseline (Pose)	2s-AGCN [20]	SH	64s	3.5M	57.1
SoA	VPN	SH	65s	24M	65.2
A (↓) or B (↓)	VPN++	SH	0.4s	14M	69.0
B (↑) or C (↓)	VPN++ + Poses	SH	64.4s	17.5M	71.0
B (↓) or C (↓)	VPN-F	SH	0.3s	12M	67.1
A (↓), C (↓), D (↑)	VPN++	NTU-120	0.35s	14M	86.7
A (↓), C (↓), D (↓)	VPN-F	NTU-60	0.28s	12M	90.8

TABLE 11: Choice of models to the practitioners. ↑ indicates High and ↓ indicates Low with (A) inference time, (B) quality of poses, (C) model size, and (D) amount of training data. Accuracy is provided for different Datasets.

8 Discussion

In this paper, we extend our previous framework Video-Pose network (VPN) to explore new mechanisms for combining video and Poses in order to classify action. Consequently, we have proposed two levels of distillation that can be adapted to different real-world application settings for recognizing actions. We summarize in table 11, the appropriate choice of distillation model or fusion mechanisms that could be exploited based on the requirements of a practitioner. Along with providing the appropriate choice of models, we also present the inference time, number of parameters of the resultant model, and action classification accuracy on relevant datasets. The choice of a model is based on factors (i.e. application requirements or network settings) concerning its applicability like (A) inference time, (B) quality of poses, (C) model size, and (D) amount of training data. From this experimental analysis, we conclude that our variants of distillation model (i.e., VPN++ and VPN-F) are useful when the end-user wants real-time predictions (e.g., low inference time), whereas the late fusion of VPN++ and Poses is preferred for offline action recognition. We notice that VPN-F is an effective model if further speed-up is required compared to VPN++ under the constraints of bad quality of Poses or less available training data. Interestingly, these lighter models are more accurate than models with similar training modalities [14], [15], [28], [30], [76].

9 Scope beyond RGB and Poses

In this section, we go beyond utilizing video-pose embedding by combining video with other modalities like Optical Flow through distillation. To this end, we investigate the applicability of using the distillation mechanisms involved in VPN++ for combining RGB and Optical Flow (OF). In our experiments, following the attempts towards distillation at feature space level [50], we use supervised contrastive distillation loss (SCD) between the features of RGB and OF streams. We dub this new Flow Augmented RGB stream as VFN++. Note that the Flow backbone for this experiment is an I3D flow stream. Experimentally, we find that attention level distillation is not effective while using optical flow as a Teacher. This might be due to high dimensionality of the flow features that hampers the attention network

Stream	SH (CS)	NTU-60 (CS)	NTU-60 (CV)
RGB	53.4	85.5	87.3
OF	51.8	85.7	92.8
RGB + OF	57.3	87.1	93.6
MARS + RGB [50]	58.1	88.2	92.9
VFN++	59.0	90.1	93.4
VFN++ + OF	**66.4**	**94.6**	**97.2**

TABLE 12: Effectiveness of Video-Flow Network++ (VFN++) representation using our SCD loss.

Fusion	SH (CS)	NTU-60 (CS)	NTU-60 (CV)
RGB + OF + 3D Poses	64.4	90.2	95.9
VPN++	69.0	91.9	94.9
VPFN++	69.7	92.1	95.5
VPFN++ + 3D Poses	71.7	95.1	98.2
VPFN++ + 3D Poses + OF	**72.9**	**96.7**	**99.1**

TABLE 13: Combination of RGB, 3D Poses and Flow modalities into a single model. Here VPFN++ is VPN++ + VFN++.

to learn relevant attention weights. Moreover, the flow features are not view-adaptive and do not consider the human anatomy while learning the attention weights. We present a comparative study with VFN++ in Table 12. We observe that VFN++ outperforms MARS+RGB due to the supervised contrastive learning mechanism. On availability of OF at inference time, the performance shoots up significantly. However this accuracy is lower than the one obtained with 3D Poses, substantiating that 3D Poses are superior than OF for ADL with subtle actions. In Table 13, we take a step forward towards combining VPN++ and VFN++. We call this resultant model Video-Pose-Flow Network++ (VPFN++). The combination is performed by the late fusion of VPN++ and VFN++ prediction scores. The minor performance improvement (+0.7% for Smarthome & +0.4% for NTU-60) in VPFN++ compared to our distillation model (VPN++) is attributed to OF distillation. So, for ADL, OF does not contribute much when 3D Poses are already well infused in RGB. With availability of Poses and OF at test time, VPFN++ + 3D Poses + OF supersedes the SoA models. Thus, our proposed framework could be extended for combining privileged modalities which is a possible perspective of this work. However, it is to be introspected that the appropriate distillation mechanism may depend on the given modalities.

10 Conclusion

In this paper, we have extended our proposed video-pose embedding for video understanding and presented a different perspective for combining RGB and 3D Poses through knowledge distillation. In an attempt to rethink combining RGB and Poses via feature fusion and attention mechanism, we propose two levels of distillation by infusing Poses at training time - feature-level and attention-level. Consequently, VPN++ does not rely anymore on the availability of 3D poses at inference time resulting in high speed up and high resiliency to noisy Poses. In addition to this, VPN++ learns a discriminative representation for classifying ADL. We show that VPN++ when combined with 3D Poses, if available, outperforms the state-of-the-art methods on 4 ADL datasets. Then, we study different strategies of combining modalities for video understanding which could be exploited by the community based on their needs.

Preliminary results show also that VPN++ can be extended to optical flow. Future work will explore towards an end-to-end framework, infusing several modalities simultaneously into a RGB stream.

Acknowledgement

We are grateful to INRIA Sophia Antipolis - Mediterranean "NEF" computation cluster for providing resources and support.

References

[1] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[2] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016, pp. 1933–1941.

[3] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4724–4733.

[4] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012. [Online]. Available: http://arxiv.org/abs/1212.0402

[5] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[6] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2556–2563.

[7] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[8] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

[10] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[11] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Val Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *ECCV*, 2016.

[12] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," *CoRR*, vol. abs/1812.02707, 2018. [Online]. Available: http://arxiv.org/abs/1812.02707

[13] N. Hussein, E. Gavves, and A. W. M. Smeulders, "Timeception for complex action recognition," *CoRR*, vol. abs/1812.01289, 2018. [Online]. Available: http://arxiv.org/abs/1812.01289

[14] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, "Glimpse clouds: Human activity recognition from unstructured feature points," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[15] S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond, and G. Francesca, "Toyota smarthome: Real-world activities of daily living," in *ICCV*, 2019.

[16] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

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

[19] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] S. Lei, Z. Yifan, C. Jian, and L. Hanqing, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019.

[21] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 9532–9545, 2020.

[22] A. Shahroudy, G. Wang, and T.-T. Ng, "Multi-modal feature fusion for action recognition in rgb-d sequences," in *2014 6th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, May 2014, pp. 1–4.

[23] G. Liu, J. Qian, F. Wen, X. Zhu, R. Ying, and P. Liu, "Action recognition based on 3d skeleton and rgb frame fusion," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov 2019, pp. 258–264.

[24] H. Rahmani and M. Bennamoun, "Learning action recognition model from depth and skeleton videos," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 5833–5842.

[25] J.-M. Perez-Rua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "Mfas: Multimodal fusion architecture search," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[26] F. Baradel, C. Wolf, and J. Mille, "Human activity recognition with pose-driven attention to rgb," in *The British Machine Vision Conference (BMVC)*, September 2018.

[27] F. Baradel, C. Wolf, and J. Mille, "Human action recognition: Pose-based attention draws focus to hands," in *proceedings of the IEEE International Conference on Computer Vision WorkshopS*, 2017, pp. 604–613.

[28] S. Das, A. Chaudhary, F. Bremond, and M. Thonnat, "Where to focus on for human action recognition?" in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Jan 2019, pp. 71–80.

[29] Z. Luo, J.-T. Hsieh, L. Jiang, J. Carlos Niebles, and L. Fei-Fei, "Graph distillation for action detection with privileged modalities," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[30] S. Das, S. Sharma, R. Dai, F. Bremond, and M. Thonnat, "Vpn: Learning video-pose embedding for activities of daily living," 2020.

[31] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal networks hard?" *CoRR*, vol. abs/1905.12681, 2019. [Online]. Available: http://arxiv.org/abs/1905.12681

[32] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 4489–4497. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.510

[33] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 5534–5542.

[34] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[35] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video classification with channel-separated convolutional networks," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[36] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," 2020.

[37] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer lstm networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2017, pp. 148–157.

[38] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 816–833.

[39] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[40] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 4263–4270.

[41] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Nov 2015, pp. 579–583.

[42] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346 – 362, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320317300936

[43] M. S. Ryoo, A. Piergiovanni, M. Tan, and A. Angelova, "Assemblenet: Searching for multi-stream neural connectivity in video architectures," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SJgMK64Ywr

[44] M. S. Ryoo, A. Piergiovanni, J. Kangaspunta, and A. Angelova, "Assemblenet++: Assembling modality representations via attention connections," in *ECCV*, 2020.

[45] S. Das, M. Thonnat, kaustubh Sakhalkar, M. Koperski, F. Brémond, and G. Francesca, "A new hybrid architecture for human activity recognition from rgb-d videos," *MultiMedia Modeling. MMM 2019*, 2019.

[46] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 892–900.

[47] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 826–834.

[48] J. Hoffman, S. Gupta, J. Leong, S. Guadarrama, and T. Darrell, "Cross-modal adaptation for rgb-d detection," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 5032–5039.

[49] N. C. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 106–121.

[50] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "MARS: Motion-Augmented RGB Stream for Action Recognition," in *CVPR*, 2019.

[51] A. Piergiovanni, A. Angelova, and M. S. Ryoo, "Evolving losses for unsupervised video representation learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[52] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *International Conference on Learning Representations*, 2020.

[53] Q. Guo, X. Wang, Y. Wu, Z. Yu, D. Liang, X. Hu, and P. Luo, "Online knowledge distillation via collaborative learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[54] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*, 2011.

[55] A. Miech, I. Laptev, and J. Sivic, "Learning a text-video embedding from incomplete and heterogeneous data," *CoRR*, vol. abs/1804.02516, 2018. [Online]. Available: http://arxiv.org/abs/1804.02516

[56] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[57] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," 2018.

[58] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-End Learning of Visual Representations from Uncurated Instructional Videos," in *CVPR*, 2020.

[59] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning, and recognition," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 2649–2656.

[60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[61] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2627435.2670313

[62] D. Yang, R. Dai, Y. Wang, R. Mallick, L. Minciullo, G. Francesca, and F. Bremond, "Selective spatio-temporal aggregation based pose refinement system: Towards understanding human activities in real-world videos," 2020.

[63] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[64] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017.

[65] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[66] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action Recognition by Dense Trajectories," in *IEEE Conference on Computer Vision & Pattern Recognition*, Colorado Springs, United States, Jun. 2011, pp. 3169–3176. [Online]. Available: http://hal.inria.fr/inria-00583818/en

[67] B. Mahasseni and S. Todorovic, "Regularizing long short term memory with 3d human-skeleton sequences for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3054–3062.

[68] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 143–152.

[69] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[70] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3671–3680.

[71] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3d action recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2842–2855, June 2018.

[72] H. Rahmani and A. Mian., "3d action recognition from novel viewpoints," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 1506–1515.

[73] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[74] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[75] H. Rahmani and A. Mian, "Learning a non-linear knowledge transfer model for cross-view action recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 2458–2466.

[76] S. Das, M. Thonnat, and F. Bremond, "Looking deeper into time for activities of daily living recognition," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 487–496.