






On the Utility of 3D Hand Poses for Action Recognition

Md Salman Shamil¹, Dibyadip Chatterjee¹, Fadime Sener²,
Shugao Ma², and Angela Yao¹

¹ National University of Singapore

² Meta Reality Labs

{salman, dibyadip, ayao}@comp.nus.edu.sg

{famesener, shugao}@meta.com

<https://s-shamil.github.io/HandFormer/>

Abstract. 3D hand pose is an underexplored modality for action recognition. Poses are compact yet informative and can greatly benefit applications with limited compute budgets. However, poses alone offer an incomplete understanding of actions, as they cannot fully capture objects and environments with which humans interact. We propose HandFormer, a novel multimodal transformer, to efficiently model hand-object interactions. HandFormer combines 3D hand poses at a high temporal resolution for fine-grained motion modeling with sparsely sampled RGB frames for encoding scene semantics. Observing the unique characteristics of hand poses, we temporally factorize hand modeling and represent each joint by its short-term trajectories. This factorized pose representation combined with sparse RGB samples is remarkably efficient and highly accurate. Unimodal HandFormer with only hand poses outperforms existing skeleton-based methods at $5\times$ fewer FLOPs. With RGB, we achieve new state-of-the-art performance on Assembly101 and H2O with significant improvements in egocentric action recognition.

Keywords: Skeleton-based action recognition · 3D hand poses · Multimodal transformer

1 Introduction

The popularity of AR/VR headsets has driven interest in recognizing hand-object interactions, particularly through egocentric [14, 26] and multiview cameras [34, 51]. Such interactions are inherently fine-grained; recognizing them requires distinguishing subtle motions and object state changes. State-of-the-art methods for hand action recognition [23, 45, 47, 66] primarily rely on multi- or single-view RGB streams, which are computationally heavy and unsuitable for resource-constrained scenarios like AR/VR. Motivated by advancements in lightweight hand pose estimation methods leveraging monochrome cameras [27, 28, 46], and the evolution of low-dimensional sensors [36, 42] such as accelerometers, MMG, EMG, demonstrating real-time hand pose estimation, we advocate

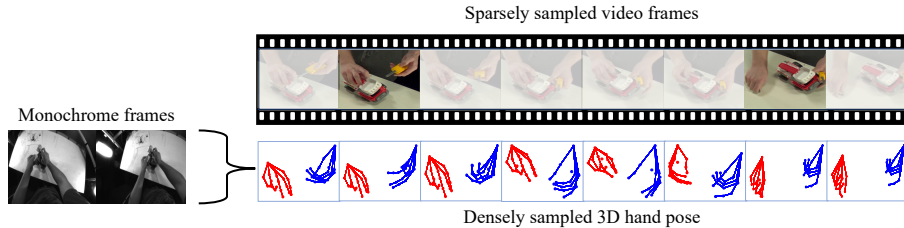


Fig. 1: We densely sample 3D hand poses to understand fine-grained hand motions and sparsely sample RGB frames to capture the scene semantics. Our 3D hand poses are acquired from low-resolution monochrome cameras but can also come from wearable sensors, facilitating an efficient understanding of hand-object interactions. Video frames and hand poses in the figure are from Assembly101 [51].

the use of 3D hand poses as an input modality for recognizing hand-object interactions. Hand poses are a compact yet informative representation that captures the motions and nuances of hand movements.

Existing works on 3D pose-based action recognition have focused primarily on full-body skeletons [20, 44, 65, 68]. Hand poses differ fundamentally from full-body skeletons. In full-body recognition datasets [41, 52], the actions are predominantly static from a global perspective. The relative changes in joint or limb positions signify the action category, *e.g.*, ‘sitting down’. Conversely, hand joints typically move together for many actions and lack a static joint as a global reference [51], *e.g.*, ‘put down toy’. Full-body skeleton methods also benefit from modeling long-range spatiotemporal dependencies between joints [44], while this is less important for the hands as shown in Fig. 2 and in Sec. 3.

However, the 3D pose alone cannot encode the hands’ actions. Unlike the full-body case, where actions are self-contained by the sequence of poses, the hands are often manipulating objects [34, 46, 51]. Hand pose is excellent for identifying motions (verbs) but struggles with associated objects [51]. Therefore, supplementing pose data with visual context from images or videos is crucial for full semantic understanding. However, as noted earlier, using dense RGB frames contradicts our motivation for using hand poses.

This paper introduces HandFormer, a novel and lightweight multimodal transformer that leverages dense 3D hand poses complemented with sparsely sampled RGB frames. To this end, we conceptualize an action as a sequence of short segments, termed *micro-actions*, which are analogous to words that form sentence-level complex actions. Each micro-action comprises a dense sequence of pose frames and a single RGB frame. As every hand joint moves in close spatial proximity, we encode the pose sequence from a Lagrangian view [49] and track each joint as an individual entity. HandFormer effectively models each micro-action whose features are then temporally aggregated for action classification. This formulation benefits from our observations on the spatiotemporal dynamics of hand movement and delivers strong performance. Additionally, limiting long-range spatiotemporal dependencies among joints and only using sparse RGB frame

features improves computational efficiency, making it well-suited for always-on, low-compute AR/VR applications.

Our contributions are: *(i)* We analyze the differences between hand pose and full-body skeleton actions and design a novel pose sequence encoding that reflects hand-specific properties. *(ii)* We propose HandFormer that takes a sequence of dense 3D hand pose and sparse RGB frames as input in the form of micro-actions. *(iii)* HandFormer achieves state-of-the-art action recognition on H2O [34] and Assembly101 [51]. Unimodal HandFormer with only hand poses outperforms existing skeleton-based methods in Assembly101, incurring at least $5\times$ fewer FLOPs. *(iv)* We experimentally demonstrate how hand poses are crucial for multiview and egocentric action recognition.

2 Related Work

Video Action recognition. Video-based action recognition systems are well-developed with sophisticated 3D-CNN [9, 23, 57, 67] or Video Transformer [2, 3, 43, 47] architectures. However, they all bear significant computational expense for both feature extraction and motion modeling, either explicitly in the form of optical flow [9, 33, 55] or implicitly through the architecture [2, 47]. Such designs are well-suited for high-fidelity applications but are not suitable for integration into lightweight systems. To this end, efficient video understanding is an active topic of research, with a focus on reducing expensive 3D operations [22, 40, 58, 67], quadratic complexity of attention [32, 43, 47] and dropping tokens [4, 21]. Additionally, *densly-sampled* RGB frames incur a high cost, yet limiting the temporal resolution will hinder fine-grained action understanding. We propose complementing 3D hand poses with *sparsely-sampled* RGB frames, developing a lightweight video understanding system.

Skeleton-based Action Recognition. Skeleton-based action recognition has been tackled with hand-crafted features [60, 62], sequence models like RNNs and LSTMs [19, 69], and CNN-based methods that either employ temporal convolution on the pose sequence [56] or transform the skeleton data into pseudo-images to be processed with 2D or 3D convolutional networks [7, 20, 31]. Recent progress has been driven by GNN-based methods that exploit the graph structure of skeletal data to construct spatio-temporal graphs and perform graph convolutions [68, 70]. These methods often model functional links between joints that go beyond skeletal connectivity [39, 54] or expand the spatiotemporal receptive field [44]. Self-attention and transformer-based methods have also been proposed [48, 65, 71]. However, most existing methods are tailored for datasets involving full-body poses. Few works [24, 30, 37, 50] study hand poses but are restricted to simple gestures, recognizable without complex temporal modeling.

Fusing RGB with Skeleton. Skeletal data has been used for cropping images of body parts [12], for weighting RGB patches around regions of interest [5, 6], or for pooling CNN features [1, 8]. Projecting into common embedding space is done in [16], enabling pose distillation [15]. Multi-stream architectures are also designed with separate paths for RGB and skeleton having lateral connec-

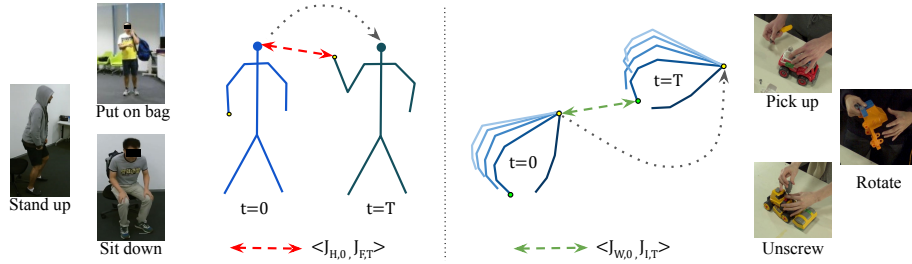


Fig. 2: Comparing skeletal changes in full-body actions from NTU RGB+D 120 [41] (left) and hand actions from Assembly101 [51] (right). Two pose frames at interval T are shown. $J_{j,t}$ indicates the 3D coordinate of joint j at timestep t , and $\langle J_x, J_y \rangle$ is the correlation between two such joints. Modeling the correlation between spatiotemporally distant joints can be informative for full-body poses but does not provide a useful action cue for hand poses.

tions between them [38, 63]. RGBPoseConv3D [20] is a two-stream model that employs 3D CNNs for both RGB and pose. Similar to skeleton-based action recognition, these multimodal approaches primarily focus on full-body poses. Some approaches simultaneously perform both hand pose estimation and action recognition, using pose data to supervise the training process [13, 64]. However, it is important to note that pose estimation deals with lower-level semantics than action recognition. Actions can often be inferred even when the estimated poses are not accurate [46].

3 Modeling Full-body vs. Hand Skeletons

Existing skeleton-based action recognition datasets primarily consist of full-body poses where actions feature significant changes over time in limb positions relative to other body parts, which highly correlate to the action category. These changes can be captured through long-range spatiotemporal modeling using graph convolutions with large receptive fields [44] or self-attention [65]. Conversely, hand actions show diverse movement patterns, as hands can move in arbitrary directions, often without significant changes in articulation. We analyze pose sequences from NTU RGB+D 120 [41] and Assembly101 [51], calculating the distance covered by each joint and identifying the least and the most active joints. We observe that, for full-body poses, these two representative joints show a significant difference in the distances covered, while for hand poses, the difference is subtle. The Pearson correlation coefficient is 0.93 for hand poses (indicating highly coupled joints) and 0.33 for full-body poses. Please refer to Suppl. Sec. A for details regarding these computations and comparisons.

An example showing the difference between body and hand poses is provided in Fig. 2, where we depict two frames separated by an interval of T frames. For the full body skeleton, the head joint J_H remains static, while the right hand joint J_F moves upward. Conversely, for the hand, all the joints move,

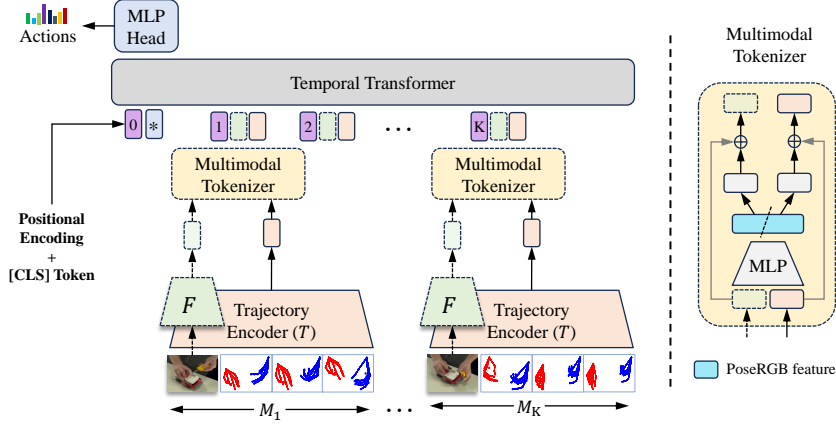


Fig. 3: Overall architecture of HandFormer. An action segment is divided into K micro-actions $\{M_1, M_2 \dots M_K\}$. Each micro-action comprises a dense sequence of pose frames and a single RGB frame. The frame encoder F and the trajectory encoder T encode the RGB and the dense poses, respectively, after which they are passed to a Multimodal Tokenizer. The modality-mixed tokens are then fed to a Temporal Transformer. Dotted paths are optional and only required when RGB is used.

including the wrist joint J_W and index fingertip J_I . The spatiotemporal correlation between the head joint in the first frame and the hand joint in the last frame ($\langle J_{H,0}, J_{F,T} \rangle$) is a crucial action cue that captures the relative structural change. However, such a correlation between the wrist and the fingertip for hand skeleton ($\langle J_{W,0}, J_{I,T} \rangle$) does not provide a stronger cue than the wrist movement itself. To avoid such redundant spatiotemporally distant correlations, we propose to divide the hand pose sequence into micro-action blocks of fixed temporal length. This formulation allows for encoding short-term movements while enabling parameter sharing.

Moreover, full-skeletal motion is a dominant feature in hand pose sequences, differentiating them from full-body poses. While the location and orientation of the human body can be trivial for most actions, this is not the case for hands. The hands do not conform to any particular 6D pose, and notably, they frequently and unpredictably alter their 6D poses over time. A comprehensive analysis of this phenomenon is provided in Suppl. Sec. A. To this end, we explicitly consider the global 6D poses of the hands during micro-action-based pose encoding.

4 HandFormer

Fig. 3 illustrates our proposed HandFormer, which consists of a sequence of micro-action blocks, a novel trajectory encoder, and a temporal aggregation module. The design of our HandFormer allows us to easily incorporate semantic context by sampling a single RGB frame from certain micro-actions.

4.1 Micro-actions

Given an action segment containing \mathcal{T} frames sampled at a certain fps, the input to our model comprises — *i*) a dense sequence of 3D hand poses, represented as $\mathbf{S} = \{P_1, P_2, \dots, P_{\mathcal{T}}\}$, where $P_t = \{P_t^{left}, P_t^{right}\} \in \mathbb{R}^{2 \times J \times 3}$ signifying the 3D coordinates of J keypoints in the left and right hands, respectively, and *ii*) a sparse set of RGB frames sampled at intervals Δf , denoted as $\mathbf{V} = \{I_1, I_{1+\Delta f}, \dots, I_{\mathcal{T}_r}\}$, where $I_t \in \mathbb{R}^{H \times W \times 3}$ are frame-wise RGB features and $\mathcal{T}_r = 1 + \left\lfloor \frac{\mathcal{T}-1}{\Delta f} \right\rfloor \times \Delta f$.

We factorize the raw input into a sequence of K micro-action blocks of length N frames, obtained by shifting the window across the action segment with a stride of R frames. Each block consists of two components — the initial appearance derived from the first RGB frame within the block and the hand motion characterized by the dense sequence of N pose frames. To obtain a fixed length input containing $\mathcal{T}' = (K-1) \times R + N$ pose frames, we perform linear interpolation for each joint along the temporal axis, transforming pose sequence \mathbf{S} having \mathcal{T} frames to $\mathbf{S}' = \{P'_1, P'_2, \dots, P'_{\mathcal{T}'}\}$ having \mathcal{T}' frames, where $P'_t \in \mathbb{R}^{2 \times J \times 3}$. Thus, we represent the input as a sequence of micro-actions $\mathbf{M} = \{M_1, M_2, \dots, M_K\}$, derived from \mathbf{V} and \mathbf{S}' through the following equations:

$$M_k = [M_k^{\text{RGB}}, M_k^{\text{Pose}}] = [I_{h(k)}, \{P'_{g(k)+i}\}_{i=0}^{N-1}], \quad (1)$$

where $g(k) = (k-1) \times R + 1$ denotes the first pose frame in k -th Micro-action, and $h(k)$ determines the index of the nearest available RGB frame.

The dense pose sequence in a micro-action captures fine-grained hand motion crucial for recognizing verbs, whereas a single RGB frame provides semantic context for recognizing objects. To extract features from micro-actions, we use a frame encoder F and a trajectory encoder T , which operate on RGB frames and pose sequences, respectively. Consequently, the RGB and pose features for the k^{th} micro-action is given by-

$$[f_k^{\text{RGB}}, f_k^{\text{Pose}}] = [F(M_k^{\text{RGB}}), T(M_k^{\text{Pose}})], \quad (2)$$

where $f_k^{\text{RGB}}, f_k^{\text{Pose}} \in \mathbb{R}^d$, where d denotes the common dimensionality of both RGB and pose embedding space.

4.2 Trajectory Encoder

We devise a trajectory-based pose encoder to encode dense hand pose sequences within micro-actions, illustrated in Fig. 4. Each joint is represented by its trajectory of dimension $3 \times N$, encapsulating the sequence of 3D coordinates across the N pose frames of a micro-action. This yields $2 \times J$ feature vectors for the J joints of two hands. Each joint's trajectory is passed through a TCN [35], whose parameters are shared for all joints. This produces $2 \times J$ Local Trajectory Tokens. Additionally, the full-skeletal motion of the hand during the entire action is used as a reference through an additional token named Global Wrist Token. This token is generated by a separate TCN operating on the sequence of 6D

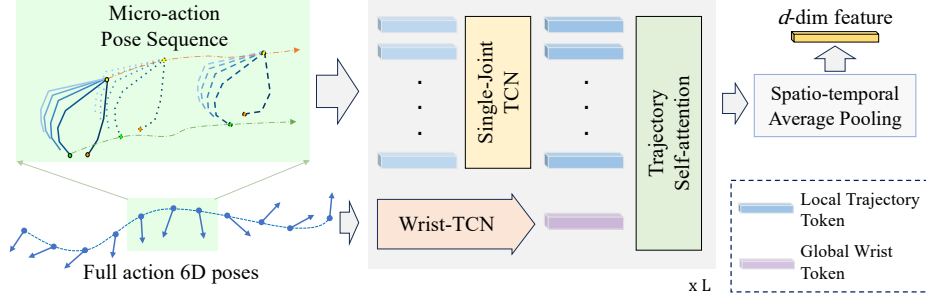


Fig. 4: Our Trajectory Encoder T , which operates on micro-actions, derives tokens with trajectory-based features and performs self-attention to encode the pose sequence into a feature vector. The Single-Joint TCN is a Temporal Convolutional Network [35] that processes the trajectories of all the joints individually with shared parameters across all joints. Wrist-TCN takes an action-wide sequence of wrist location and hand orientation (6D pose) to produce a global reference token.

poses of hands, indicating wrist location and hand orientation. Subsequently, a self-attention layer is applied to these trajectory tokens, preserving the temporal dimension for subsequent stages. This iterative process culminates in spatiotemporal average pooling, summarizing the hand motion of the micro-action.

4.3 Multimodal Tokenizer

This section discusses incorporating sparsely sampled RGB frames into HandFormer to better capture scene semantics. A single frame is sampled from each micro-action, generating an extended crop ($1.25\times$) around the hands. While the full image provides an overall scene context, the crop focuses specifically on hand-object interaction regions [10]. Features for both are separately generated using a pre-existing image encoder and then aggregated to enrich the hand-object interaction feature with scene context. The hand-object ROI crop can be obtained using an off-the-shelf HOI detector [53] or by using the corresponding hand pose projections. We opt for the latter.

Our multimodal tokenizer in Fig. 3 generates RGB and Pose tokens enhanced via multimodal interactions. This involves concatenating each frame feature and trajectory encoding, and projecting them into a shared PoseRGB feature space using an MLP. This PoseRGB feature is then split into two parts, which are added to the original frame feature and trajectory encoding, respectively.

4.4 Temporal Transformer

The multimodal tokenizer provides RGB and pose tokens for each micro-action. Since an action segment consists of a sequence of micro-actions, these micro-action tokens are aggregated over time via a temporal transformer. A video \mathbf{V} ,

divided into K micro-actions can be represented by two sets of tokens $\{\hat{f}_k^{\text{RGB}}\}_{k=1}^K$ and $\{\hat{f}_k^{\text{Pose}}\}_{k=1}^K$, respectively, produced by the multimodal tokenizer. These $2 \times K$ tokens of dimension d form the input sequence of the temporal transformer. Positional encoding and modality embedding are added to each input token to indicate the temporal location and source modality. We use the fixed sine/cosine positional encoding [59], assigning the same position to two tokens from the same micro-action. Modality embeddings are learned and shared across the tokens from the same modality (RGB or Pose). Following standard practice [17, 18], we prepend an additional learnable class token $[\text{CLS}] \in \mathbb{R}^d$, the output of which is then used for classifying actions.

4.5 Learning Objectives

HandFormer is trained end-to-end for action recognition, supervised by a cross-entropy loss: $\mathcal{L}_{cls} = -a_i \log \hat{a}_i$, where a_i represents the ground truth action label for the i^{th} sample, and \hat{a}_i denotes the predicted action category. The input pose and RGB modalities provide complementary information regarding a scene, capturing motion and interacting objects, respectively. To effectively utilize this, we employ explicit verb and object supervision (\mathcal{L}_{verb} , \mathcal{L}_{obj}) via two additional learnable class tokens. Given that pose strongly correlates with verbs, whereas objects are identifiable from RGB frames alone, we let the verb class token attend exclusively to pose encodings and the object class token to frame features.

Feature Anticipation Loss. Hand pose sequence captures the primary sources of state changes during hand-object interactions. We posit that the visual state from an initial RGB frame, in combination with the subsequent hand pose sequence, is indicative of the visual state that results from the completion of the sequence. Therefore, given an RGB feature and the corresponding pose features from a micro-action, we force our model to anticipate the RGB feature for the next micro-action by minimizing an L_1 feature loss. This loss, inspired by existing efforts [25, 61], quantifies the difference between the *anticipated* image feature and the true feature extracted from a frozen image encoder. Formally,

$$\mathcal{L}_{ant} = \sum_{k=1}^{K-1} \|\Phi_{\text{ant}}(f_k^{\text{PoseRGB}}) - f_{k+1}^{\text{RGB}}\|_1 \quad (3)$$

where Φ_{ant} denotes a linear projection layer. Hence, the total loss is-

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{verb} + \lambda_2 \mathcal{L}_{obj} + \lambda_3 \mathcal{L}_{ant} \quad (4)$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters used to balance the four losses.

5 Experiments

Datasets. We conduct experiments on two publicly available hand-object interaction datasets. **Assembly101** [51] is a large-scale multiview dataset that fea-

tures videos of procedural activities for assembling and disassembling 101 take-apart toy vehicles. This dataset has 1380 fine-grained actions based on 24 verbs and 90 objects. It consists of 12 temporally synchronized views — eight static and four egocentric. The egocentric views exhibit low-resolution monochrome images, making it challenging even for human eyes to discern objects in view. Hence, we opt for fixed views as our RGB modality. 3D hand poses estimated using off-the-shelf MEgATrack [27] are provided by the dataset. **H2O** [34] features four participants performing 36 actions that involve 8 objects and 11 verbs. Ground truth 3D hand poses are provided. The dataset consists of 5 temporally synchronized RGB views — 4 fixed and 1 egocentric. We use RGB frames from the egocentric view only.

Implementation Details. Similar to [65], our pose input consists of 120 frames by setting $\mathcal{T}' = 120$. The number of frames per micro-action (N) is 15. For multimodal experiments, we set the window stride $R = N$ and take $K = 8$ non-overlapping micro-actions, which allows us to sample 8 RGB frames for each video following [51]. However, we allow a 50% overlap between consecutive micro-actions for pose-only variants. A frozen ViT [18] or ResNet [29], followed by a learnable linear layer, is used as the frame encoder F . More details are provided in Suppl. Sec. C. Each model is trained for 50 epochs using SGD with momentum 0.9, a batch size of 32, a learning rate of 0.025, and step LR decay with a factor of 0.1 at epochs $\{25, 40\}$. Loss hyperparameters $\{\lambda_1, \lambda_2, \lambda_3\}$ are chosen to be $\{1.0, 1.0, 2.0\}$.

Model Variants. We propose several variants of our model by adjusting the width d and the number of layers T_n of the transformer to balance efficiency and accuracy. Our default HandFormer, denoted HandFormer-B, has parameters $(d, T_n) = (256, 2)$. We introduce a larger variant, HandFormer-L, with $(d, T_n) = (512, 4)$. We also explore different configurations for the number of input joints J per hand. Unless otherwise mentioned, we utilize all 21 joints per hand along with the base model denoted as HandFormer-B/21 while offering a highly efficient option utilizing only six joints per hand (five fingertips and the wrist) termed HandFormer-B/6. Additionally, HandFormer-X/ $J \times \mathcal{N}$ denotes a multimodal variant that utilizes \mathcal{N} RGB frames, where $\mathcal{N} \leq K$.

5.1 Comparison with State-of-the-Art

To evaluate the effectiveness of our proposed architecture, we compare it against several baselines. For pose-only comparisons, we choose a graph-based network MS-G3D [44] and an attention-based method ISTA-Net [65] — the two best performing skeleton-based methods on Assembly101 [51], as reported by [65]. We also employ video baselines TSM [40] and SlowFast [23], emphasizing efficiency and high temporal resolution, respectively. Additionally, we include H2OTR [13], the state-of-the-art for the H2O dataset. For Assembly101, we replicate the results of the methods above using RGB frames from *view 4* of the dataset, while results on H2O are acquired from the respective papers. Furthermore, on both datasets, we train and test RGBPoseConv3D [20], which is current state-of-the-art in multimodal action recognition with skeleton and RGB data.

Table 1: Quantitative comparison with state-of-the-art methods on Assembly101 [51] and H2O [34]. For H2O, 6D object pose is used by ISTA-Net during training and inference and by H2OTR during training. ‘MS-G3D + TSM’ denotes a late fusion of the corresponding unimodal architectures. Reported GFLOPs include pose estimation costs in the context of Assembly101 at 15 fps. [†]For pose-only methods, pose estimation costs are common and excluded for simpler comparison.

Method	Pose	RGB	GFLOPs	Assembly101			H2O
				Action	Verb	Object	Action
MS-G3D [44]	✓	✗	21.2 [†]	28.78	63.46	37.26	50.83
ISTA-Net [65]	✓	✗	35.2 [†]	28.14	62.70	36.77	89.09
SlowFast [23]	✗	✓	65.7	-	-	-	77.69
TSM [40]	✗	✓	33.0	35.27	58.27	47.45	-
H2OTR [13]	✗	✓	-	-	-	-	90.90
RGBPoseConv3D [20]	✓	✓	68.9	33.61	61.99	42.90	83.47
MS-G3D + TSM	✓	✓	66.2	39.74	65.12	51.12	-
HandFormer-B/21×8	✓	✗	4.2 [†]	28.80	65.33	36.28	57.44
	✗	✓	33.0	32.07	55.61	44.89	84.71
	✓	✓	47.6	41.06	69.23	51.17	93.39

We evaluate three variants of our method by controlling the input modalities. As shown in Tab. 1, our unimodal pose-only model excels in verb recognition on Assembly101. In contrast, RGB-based methods benefit from object appearance and usually perform well for actions due to the strong object recognition.

In this context, the accuracy of ISTA-Net on H2O is not directly comparable to our method, as they also use the 6D object poses, while we only use hand poses as input. RGBPoseConv3D struggles to achieve satisfactory performance, particularly on Assembly101, suggesting that generalizing to hand poses is non-trivial for skeleton-based methods. Therefore, we combine two best-performing unimodal methods with late fusion (MS-G3D + TSM) to create a stronger baseline. Our model even outperforms this, indicating the effectiveness of our proposed multimodal fusion.

5.2 Skeleton-based Action Recognition for Hands

The compositional nature of action classes in hand-object interaction videos allows us to break down the action into a verb and an object. Recognizing such actions from 3D hand poses is an ill-posed problem, as the hand skeletons lack explicit information about the interacting objects, which are also part of the action semantics. However, as the pose data completely captures the hand motion information, it can be reliably used for verb recognition. Therefore, we evaluate a pose-only version of our method on the verb recognition task and compare the performance and efficiency metrics with other state-of-the-art skeleton-based methods in Fig. 5. Our method uses significantly fewer GFLOPs due to our spatiotemporal factorization using micro-actions. With $J = 21$, all our HandFormer variants outperform existing methods. The $J = 6$ variant is exceptionally efficient with comparable accuracy, notably surpassing the efficiency-focused Shift-GCN [11]. HandFormer-H/21 combines the HandFormer-B and HandFormer-L

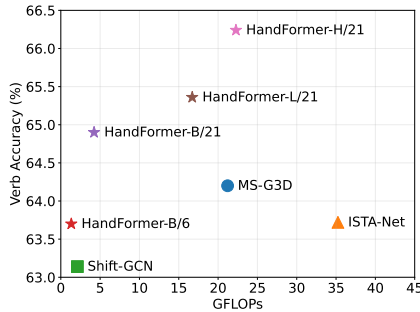


Fig. 5: Comparison of skeleton-based methods for verb recognition on Assembly101 [51]. Our method achieves state-of-the-art performance while utilizing significantly fewer FLOPs.

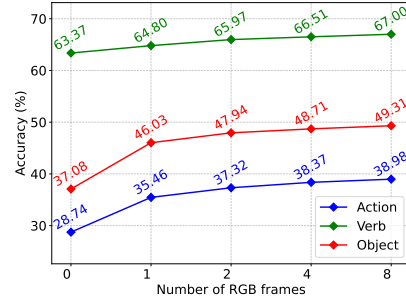


Fig. 6: Ablating number of RGB frames in HandFormer on Assembly101 [51]. With more RGB frames, verb recognition shows marginal gain, whereas object recognition shows improvement with diminishing returns.

variants and becomes our best-performing model, improving over MS-G3D by 2.04% while maintaining comparable FLOPs.

5.3 How many RGB frames are required?

In our model, the pose modality maintains a high temporal resolution to capture fine-grained hand movements, resulting in good verb recognition performance. On the contrary, RGB frames are primarily required to introduce semantic context for object recognition and does not necessitate a high temporal resolution like hand movements. The design of our model allows us to sample only a few RGB frames (as low as one) but still perform competitively at a reduced computational cost. Fig. 6 shows the impact of using more RGB frames for Assembly101 [51]. For this experiment, non-overlapping micro-actions are considered. Using only one RGB frame in HandFormer (35.46) outperforms the video model TSM (35.27), as shown in Tab. 1. This performance gain stems primarily from a significant improvement in object accuracy, with only a slight enhancement in verb accuracy. However, including more RGB frames shows a diminishing return, which is unsurprising as additional frames are expected to provide redundant information. These results are obtained with a simplified version of our model by setting $\mathcal{L} = \mathcal{L}_{cls}$ and bypassing the multimodal tokenizer.

5.4 Can 3D hand pose be an efficient alternative to multi-view?

While multiview action recognition benefits from precise hand-movement information in 3D space, processing all views with video models is expensive and highly redundant. In Tab. 2, we demonstrate that combining 3D hand pose with a single RGB view (*view 4*) achieves comparable performance to multi-view action recognition on Assembly101 [51]. Specifically, our action recognition

Table 2: Multi-view Action Recognition on Assembly101 [51]. HandFormer with ‘Single View+*Pose*’ has better performance than ‘Single View+*Egocentric*’. Also, verb performance is comparable to using all (8) RGB views.

Views	Action	Verb	Object
Single View	35.27	58.27	47.45
Single View + <i>Egocentric</i>	37.75	61.80	49.43
Two Views	41.96	65.22	53.26
All (8) Views	47.51	70.99	57.73
Single View + <i>Pose</i>	41.06	69.23	51.17

Table 3: Cross-view performance of HandFormer in Assembly101 [51] shows its generalization capability to unseen *view 1*, outperforming the video baseline which is trained on *view 1* directly. Egocentric views are the source of hand poses here and, therefore, are included in the video models.

Method	Train on	Test on	Action	Verb	Object
TSM [40]	<i>v4</i> + ego	<i>v4</i> + ego	37.75	61.80	49.43
	<i>v4</i> + ego	<i>v1</i> + ego	35.27	59.53	47.72
	<i>v1</i> + ego	<i>v1</i> + ego	36.21	60.78	48.52
Our Method	<i>v4</i> + Pose	<i>v4</i> + Pose	41.06	69.23	51.17
	<i>v4</i> + Pose	<i>v1</i> + Pose	38.43	67.86	48.32

performance (‘Single View + *Pose*’) matches the fusion of the two most informative views — *view 1* and *view 4*. Notably, our verb recognition accuracy is on par with combining all 8 RGB views, though the latter incurs at least $5\times$ more FLOPs despite using the efficient TSM model. Additionally, using hand pose in combination with an RGB view (‘Single View + *Pose*’) outperforms directly using the egocentric videos (‘Single View + *Egocentric*’) from which the hand poses are derived. While fusing multiple RGB views improves accuracy by ensembling multiple complementary predictions, the computational overhead increases significantly. In contrast, our model processes hand pose and single-view RGB frames, enhancing efficiency by leveraging the less redundant and low-dimensional pose data.

Cross-view Generalization. 3D hand pose offers a unique opportunity for cross-view generalization because of its universality across different viewpoints. To evaluate the effectiveness of our method for unseen views, we train our model with frame-wise RGB features from *view 4* and test it on *view 1*. We train TSM [40] as a baseline video model on both RGB views separately. As our method includes 3D hand poses obtained from the egocentric views, we include the egocentric videos in the TSM baselines for a fair comparison. As seen from Tab. 3, our method, trained on *view 4*, generalizes well on unseen *view 1*, outperforming the TSM model that was directly trained on *view 1*.

5.5 Egocentric Action Recognition

Recognizing actions in egocentric videos is challenging due to camera motion and the occlusion of interacting objects by the hands. Additionally, in the case of Assembly101 [51], the egocentric cameras are similar to Oculus Quest VR headsets, which provide monochrome low-resolution frames. As a result, action recognition performance is significantly lower than fixed RGB views, as found in [51]. We address this challenging scenario with our proposed multimodal architecture and achieve state-of-the-art performance in egocentric action recognition on Assem-

Table 4: Egocentric action recognition in Assembly101 [51]. TSM features from [51] are used as RGB frame features.

Method	Action	Verb	Object
TSM egocentric (fuse 4 views)	33.80	59.00	46.50
Egocentric ($e3$) + Pose	36.07	65.52	45.82
Egocentric ($e4$) + Pose	35.56	65.79	45.20
Egocentric ($e3+e4$) + Pose	38.05	66.32	47.86

Table 5: Keypoint ablation for verb recognition in Assembly101 [51]

#Joints	Global Reference	Verb Accuracy (%)
21	✗	64.17
21	✓	64.90
11	✓	64.77
6	✓	63.70

Table 6: Micro-action length ablation for verb recognition in Assembly101 [51]

#Frames	1	15	30	60	120
Verb Accuracy(%)	59.12	63.70	63.68	63.51	62.29

Table 7: Ablating unimodal temporal aggregation for verb recognition in Assembly101 [51]

Temp. Agg.	TCN	LSTM	Transformer
Verb Acc. (%)	62.95	63.34	63.70

bly101. For this experiment, we used frame-wise TSM features provided by [51]. As depicted in Tab. 4, our model, using a single egocentric view ($e3$ or $e4$) outperforms the fusion of four egocentric views as reported in [51]. Moreover, fusing $e3$ and $e4$ significantly enhances our model, resulting in a 4.25% increase in action accuracy over the baseline.

5.6 Ablation Studies

Keypoints. Not all joints of the hands are equally informative for understanding hand actions. For instance, fingertips exhibit greater mobility compared to the inner joints. Moreover, from an egocentric viewpoint, certain joints are more prone to self-occlusion than others. In Tab. 5, we present the impact of incorporating varying numbers of joints on verb recognition within the Assembly101 dataset [51]. For the case of 6 joints, we consider only the wrist joint along with the five fingertips. To expand to 11 joints, we incorporate all the joints along the index and the thumb, which are least affected by self-occlusion. We also show the effect of including the Global Wrist Token, which acts as a reference to the global motion of the hands while encoding micro-actions.

Micro-action length. As the resized input comprises a fixed number of pose frames, enlarging the window size for a micro-action decreases the number of micro-actions to aggregate, and vice-versa. In Tab. 6, we vary the micro-action length for verb recognition in the Assembly101 [51] using 6 joints per hand as input, *i.e.* fingertips and wrist joint. The input pose clip is temporally resized to $T' = 120$ before breaking into micro-actions. Lengths 1 and 120 represent two extreme versions with frame-based and trajectory-based encoding, respectively, while the others conform to our micro-action-based formulation.

Table 8: Ablating tokenization and different losses for action recognition

Multimodal Tokenizer	Feature Ant. Loss	Verb & Object Loss	Action Accuracy (%)	
			Assembly101 [51]	H2O [34]
\times	\times	\times	38.98	85.95
\checkmark	\times	\times	40.19	88.84
\checkmark	\checkmark	\times	40.24	89.26
\times	\times	\checkmark	40.56	90.50
\checkmark	\checkmark	\checkmark	41.06	93.39

Temporal Aggregation. After extracting micro-action features, aggregation for action recognition can be done with any sequence model. In Tab. 7, we evaluate the effectiveness of different temporal aggregation methods for verb recognition in the Assembly101 dataset [51]. Here, we use 6 joints per hand as input, *i.e.*, the fingertips and the wrist joint.

Loss components. To assess the individual contributions of different components, we begin with a basic configuration. We then systematically introduce each element to understand its impact on the overall performance as observed in Tab. 8. Incorporating modality interaction between RGB and pose at the micro-action level through a multimodal tokenizer enhances action accuracy. The introduction of auxiliary losses also has a positive impact, resulting in an overall improvement of 2.08% for Assembly101 [51] and 7.44% for H2O [34].

6 Conclusion

With the growing interest in AR/VR and wearables, hand pose estimation has rapidly advanced, and holds promise as a compact and domain-independent modality to complement the visual input. To address the underexplored domain of using 3D hand pose as a modality for hand-object interaction recognition, we introduce HandFormer, a novel multimodal transformer that leverages dense sequences of 3D hand poses with sparsely sampled RGB frames to achieve state-of-the-art action recognition performance. Our model also reduces computational requirements, offering immediate significance across various low-resource applications in mobile devices.

Limitations. Our method relies on the availability of hand poses, which, if extracted from the visual modality with pose estimation tools [27, 28], can encounter out-of-view scenarios and produce noisy poses. Our experiments reveal that these estimated poses can still achieve good accuracy, yet further research can be conducted to explicitly address this phenomenon. We also assume that a uniform sampling of RGB frames from each micro-action should provide good representations for understanding the semantic context. However, not all frames are equally important in understanding the action. In such a case, adaptive frame sampling methods can be employed, which we leave for future work.

Acknowledgements

This research is supported by A*STAR under its National Robotics Programme (NRP) (Award M23NBK0053).

References

1. Ahn, D., Kim, S., Hong, H., Ko, B.C.: Star-transformer: a spatio-temporal cross attention transformer for human action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3330–3339 (2023)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)
3. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)
4. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022)
5. Bruce, X., Liu, Y., Chan, K.C.: Multimodal fusion via teacher-student network for indoor action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3199–3207 (2021)
6. Bruce, X., Liu, Y., Zhang, X., Zhong, S.h., Chan, K.C.: Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(3), 3522–3538 (2022)
7. Caetano, C., Sena, J., Brémond, F., Dos Santos, J.A., Schwartz, W.R.: Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In: 2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS). pp. 1–8. IEEE (2019)
8. Cao, C., Zhang, Y., Zhang, C., Lu, H.: Body joint guided 3-d deep convolutional descriptors for action recognition. *IEEE transactions on cybernetics* **48**(3), 1095–1108 (2017)
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
10. Chatterjee, D., Sener, F., Ma, S., Yao, A.: Opening the vocabulary of egocentric actions. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
11. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 183–192 (2020)
12. Chéron, G., Laptev, I., Schmid, C.: P-cnn: Pose-based cnn features for action recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 3218–3226 (2015)
13. Cho, H., Kim, C., Kim, J., Lee, S., Ismayilzada, E., Baek, S.: Transformer-based unified recognition of two hands manipulating objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4769–4778 (2023)

14. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European conference on computer vision (ECCV). pp. 720–736 (2018)
15. Das, S., Dai, R., Yang, D., Bremond, F.: Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(12), 9703–9717 (2021)
16. Das, S., Sharma, S., Dai, R., Bremond, F., Thonnat, M.: Vpn: Learning video-pose embedding for activities of daily living. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 72–90. Springer (2020)
17. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018)
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
19. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1110–1118 (2015)
20. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2969–2978 (2022)
21. Fayyaz, M., Koohpayegani, S.A., Jafari, F.R., Sengupta, S., Joze, H.R.V., Sommerlade, E., Pirsavash, H., Gall, J.: Adaptive token sampling for efficient vision transformers. In: European Conference on Computer Vision. pp. 396–414. Springer (2022)
22. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 203–213 (2020)
23. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
24. Gan, M., Liu, J., He, Y., Chen, A., Ma, Q.: Keyframe selection via deep reinforcement learning for skeleton-based gesture recognition. *IEEE Robotics and Automation Letters* (2023)
25. Girdhar, R., Grauman, K.: Anticipative video transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13505–13515 (2021)
26. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19012 (2022)
27. Han, S., Liu, B., Cabezas, R., Twigg, C.D., Zhang, P., Petkau, J., Yu, T.H., Tai, C.J., Akbay, M., Wang, Z., et al.: Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (ToG)* **39**(4), 87–1 (2020)
28. Han, S., Wu, P.c., Zhang, Y., Liu, B., Zhang, L., Wang, Z., Si, W., Zhang, P., Cai, Y., Hodan, T., et al.: Umetrack: Unified multi-view end-to-end hand tracking for vr. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)

29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
30. Hou, J., Wang, G., Chen, X., Xue, J.H., Zhu, R., Yang, H.: Spatial-temporal attention res-tcn for skeleton-based dynamic hand gesture recognition. In: Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0–0 (2018)
31. Hou, Y., Li, Z., Wang, P., Li, W.: Skeleton optical spectra-based action recognition using convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology* **28**(3), 807–811 (2016)
32. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International conference on machine learning. pp. 4651–4664. PMLR (2021)
33. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5492–5501 (2019)
34. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10138–10148 (2021)
35. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 156–165 (2017)
36. Lee, C.J., Zhang, R., Agarwal, D., Yu, T.C., Gunda, V., Lopez, O., Kim, J., Yin, S., Deng, B., Li, K., et al.: Echowrist: Continuous hand pose tracking and hand-object interaction recognition using low-power active acoustic sensing on a wristband. arXiv preprint arXiv:2401.17409 (2024)
37. Li, C., Li, S., Gao, Y., Zhang, X., Li, W.: A two-stream neural network for pose-based hand gesture recognition. *IEEE Transactions on Cognitive and Developmental Systems* **14**(4), 1594–1603 (2021)
38. Li, J., Xie, X., Pan, Q., Cao, Y., Zhao, Z., Shi, G.: Sgm-net: Skeleton-guided multimodal network for action recognition. *Pattern Recognition* **104**, 107356 (2020)
39. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3595–3603 (2019)
40. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7083–7093 (2019)
41. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence* **42**(10), 2684–2701 (2019)
42. Liu, Y., Zhang, S., Gowda, M.: Neuropose: 3d hand pose tracking using emg wearables. In: Proceedings of the Web Conference 2021. pp. 1471–1482 (2021)
43. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)
44. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 143–152 (2020)

45. Ma, J., Damen, D.: Hand-object interaction reasoning. In: 2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–8. IEEE (2022)
46. Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., Keskin, C.: Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12999–13008 (2023)
47. Patrick, M., Campbell, D., Asano, Y., Misra, I., Metze, F., Feichtenhofer, C., Vedaldi, A., Henriques, J.F.: Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems* **34**, 12493–12506 (2021)
48. Plizzari, C., Cannici, M., Matteucci, M.: Spatial temporal transformer network for skeleton-based action recognition. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III. pp. 694–701. Springer (2021)
49. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Feichtenhofer, C., Malik, J.: On the benefits of 3d pose and tracking for human action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 640–649 (2023)
50. Sabater, A., Alonso, I., Montesano, L., Murillo, A.C.: Domain and view-point agnostic hand action recognition. *IEEE Robotics and Automation Letters* **6**(4), 7823–7830 (2021)
51. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhanian, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21096–21106 (2022)
52. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1010–1019 (2016)
53. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9869–9878 (2020)
54. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12026–12035 (2019)
55. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems* **27** (2014)
56. Soo Kim, T., Reiter, A.: Interpretable 3d human action analysis with temporal convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 20–28 (2017)
57. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
58. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
59. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)

60. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 588–595 (2014)
61. Vondrick, C., Pirsiaavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 98–106 (2016)
62. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 1290–1297. IEEE (2012)
63. Weiyao, X., Muqing, W., Min, Z., Ting, X.: Fusion of skeleton and rgb features for rgb-d human action recognition. *IEEE Sensors Journal* **21**(17), 19157–19164 (2021)
64. Wen, Y., Pan, H., Yang, L., Pan, J., Komura, T., Wang, W.: Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric rgb videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21243–21253 (2023)
65. Wen, Y., Tang, Z., Pang, Y., Ding, B., Liu, M.: Interactive spatiotemporal token attention network for skeleton-based general interactive action recognition. arXiv preprint arXiv:2307.07469 (2023)
66. Wu, C.Y., Li, Y., Mangalam, K., Fan, H., Xiong, B., Malik, J., Feichtenhofer, C.: Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13587–13597 (2022)
67. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 305–321 (2018)
68. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
69. Zhang, S., Liu, X., Xiao, J.: On geometric features for skeleton-based action recognition using multilayer lstm networks. In: 2017 IEEE winter conference on applications of computer vision (WACV). pp. 148–157. IEEE (2017)
70. Zhang, X., Xu, C., Tian, X., Tao, D.: Graph edge convolutional neural networks for skeleton-based action recognition. *IEEE transactions on neural networks and learning systems* **31**(8), 3047–3060 (2019)
71. Zhang, Y., Wu, B., Li, W., Duan, L., Gan, C.: Stst: Spatial-temporal specialized transformer for skeleton-based action recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3229–3237 (2021)