

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 Research

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

{famesener, shugao}@meta.com

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

Abstract. 3D hand poses are an under-explored 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. To efficiently model hand-object interactions, we propose HandFormer, a novel multimodal transformer. 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 achieves high accuracy. 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, 25] and multi-view cameras [33, 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, 44, 47, 67] 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 methodologies leveraging monochrome cameras [26, 27, 45], and the evolution of low-dimensional sensor technologies [35, 41] such as accelerometers, MMG, EMG, demonstrating real-time hand pose estimation, we advocate for the utilization

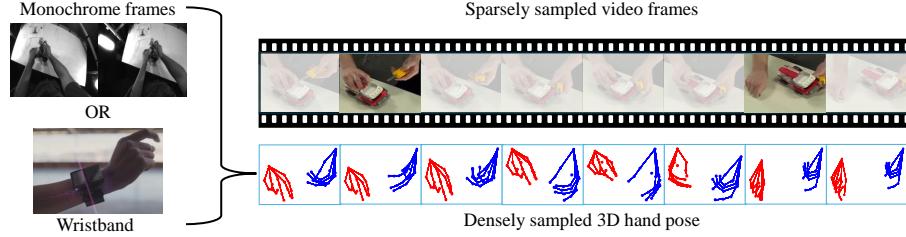


Fig. 1: We sample 3D hand poses at a high temporal resolution (dense) for understanding fine-grained hand motion and sparsely sample RGB frames to capture the scene semantics. 3D hand poses can be acquired from low-resolution monochrome cameras or low-dimensional sensors such as accelerometers, MMG, EEG, etc., facilitating an efficient understanding of hand-object interactions. Video frames and hand poses in the figure are from Assembly101 [51].

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, 43, 66, 69]. Hand poses differ fundamentally from full-body skeletons. In full-body recognition datasets [40, 53], the actions are predominantly static from a global perspective. The relative changes in joint or limb positions signify the action category, *e.g.* ‘hand waving.’ 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 [43], while this is less important for the hands as shown in Fig. 2 and in Sec. 3.

However, the 3D pose alone is insufficient to encode the action for hands. Unlike the full-body case, where actions are self-contained by the sequence of poses, the hands are often manipulating objects [33, 45, 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 we noted earlier, using dense RGB frames contradicts our objective and motivation for using hand poses.

This work 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, which we refer to as *micro-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 to each other, we opt to encode the pose sequence from a Lagrangian view [49] and track each joint as an individual entity. The pose trajectory and a single RGB frame are combined into micro-action features, which are then temporally aggregated with a transformer to predict the action label.

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 performance on H2O [33] and Assembly101 [51] while unimodal HandFormer with only hand poses outperforms existing skeleton-based methods for verb recognition in Assembly101 incurring at least 5 \times fewer FLOPs. *(iv)* We experimentally demonstrate how hand poses are especially crucial for multi-view and egocentric action recognition.

2 Related Work

Video Action recognition. Video-based action recognition systems are well-developed with sophisticated 3D-CNN [9, 23, 58, 68] or Video Transformer [2, 3, 42, 47] architectures. However, they all bear significant computational expense for both feature extraction and motion handling, either explicitly in the form of optical flow [9, 32, 56] or implicitly through the architecture [2, 47]. Such designs are well-suited for high-fidelity research and certain cloud-based applications but are not suitable for integration into lightweight systems. To this end, efficiency in video understanding has been an active topic of research where the main focus has been model efficiency i.e., reducing expensive 3D operations [22, 39, 59, 68], quadratic complexity of attention [31, 42, 47] and dropping tokens [4, 21]. However, due to the high cost of using *densely-sampled* RGB frames, they can only offer limited temporal resolution, hindering fine-grained action understanding. As such, we propose to complement 3D hand poses with only *sparsely-sampled* RGB frames for developing a lightweight video understanding system.

Skeleton-based Action Recognition. Skeleton-based action recognition has been approached with hand-crafted features [61, 63], sequence models like RNNs and LSTMs [19, 70], and CNN-based methods that either employ temporal convolution on the pose sequence [57] or transform the skeleton data into pseudo-images to be processed with 2D or 3D convolutional networks [7, 20, 30]. Recent advancements primarily come from GNN-based methods that utilize the skeletal data's graph structure for constructing spatio-temporal graphs and performing graph convolutions [69, 71]. These methods often model functional links between joints that go beyond skeletal connectivity [38, 55] or aim to increase the spatiotemporal receptive field [43]. Self-attention and transformer-based methods have also been proposed in [48, 66, 72]. Most existing methods are tailored for datasets that involve full-body poses, while our work is dedicated to 3D hand pose data, focusing on the unique motion characteristics of hand skeletons. Although there are some methods that work with hand pose, such as [29, 36, 50], they are primarily designed for gesture recognition, which does not require explicit temporal modeling.

Fusing RGB with Skeleton. Pose or skeletal data can be used for cropping image patches for body parts [12], weighting RGB patches around the regions of interest [5, 6], or pooling deep CNN features [1, 8]. Projecting into common

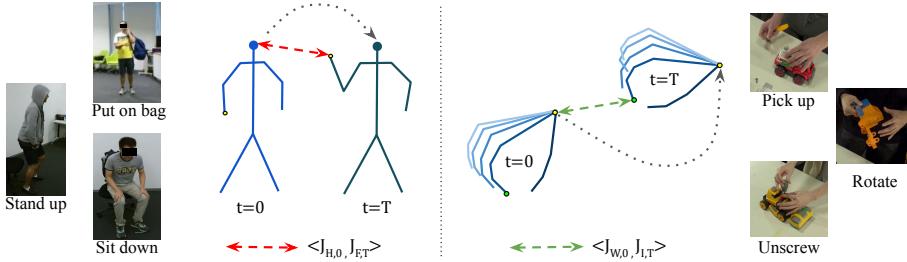


Fig. 2: Comparing skeletal changes in full-body actions from NTU RGB+D 120 [40] (left) and hand actions from Assembly101 [52] (right). Two pose frames at interval T are shown. $J_{j,t}$ indicates the 3D coordinate of joint j at timestep t , and $\langle J_x, J_y \rangle$ is the correlation between two such joints. Modeling the correlation between spatio-temporally distant joints can be informative in full-body pose but is unable to provide a useful action cue in hand pose.

embedding space is done in [16], which enables pose distillation in [15]. Multi-stream architectures are also designed to employ separate paths for RGB and skeleton data with lateral connections between the streams [37, 64]. RGBPoseConv3D [20] is a two-stream architecture that proposes to use 3D CNNs for both RGB and pose. Similar to skeleton-based action recognition, these multi-modal approaches combining skeleton and RGB data 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, 65]. However, it is important to note that the task of pose estimation is extremely low-level when compared to action recognition, which deals with high-level semantics. Actions can often be inferred even when the estimated poses are not highly precise [45].

3 Full-body Skeleton vs. Hand

Existing skeleton-based action recognition datasets primarily consist of full-body poses where actions feature significant changes in limb positions relative to other body parts over time, strongly correlating with the action category. Capturing these changes can be achieved through long-range spatiotemporal modeling using graph convolutions with large receptive fields [43] or using self-attention [66]. Conversely, hand-pose actions show diverse movement patterns, as hands can move in arbitrary directions, often without significant changes in articulation. We also analyze pose sequences from NTU RGB+D 120 [40] and Assembly101 [51], calculating the distance covered by each joint and identifying the least and most active joints. We observe that, for full-body poses, these two representative joints show a significant difference in the distance covered, while for hand poses, they only depict a subtle distinction. The Pearson correlation coefficient is 0.93 for hand poses (indicating highly coupled joints) and 0.33 for full-body poses. Please see our Supplementary for the details of these computations and comparisons.

An example showing the difference between body and hand pose is provided in Fig. 2, where we depict two frames separated by a time interval of T frames.

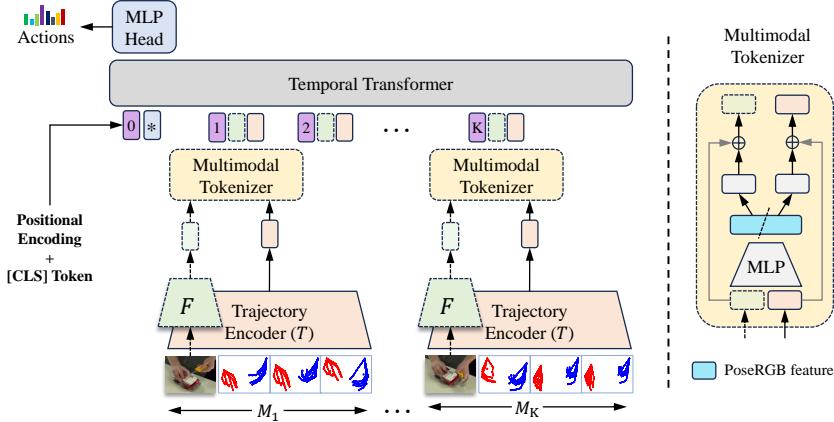


Fig. 3: Overall architecture of HandFormer. An action segment is divided into K micro-actions $\{M_1, M_2, \dots, M_K\}$. Each micro-action comprises a dense sequence of pose frames and a single RGB frame. The frame encoder F and the trajectory encoder T encode the single RGB frame and the dense poses,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

For the full body skeleton, the head joint J_H remains static, while the right hand joint J_F moves upward. Conversely, in the hand pose, all the joints move, including the wrist joint J_W and index fingertip J_I . The spatiotemporal cross-correlation between the head joint in the first frame and the hand joint in the last frame ($<J_{H,0}, J_{F,T}>$) is an important action cue, capturing the relative structural change. On the contrary, modeling a similar correlation between the wrist and a fingertip for hand skeleton ($<J_{W,0}, J_{I,T}>$) does not provide a stronger cue than the wrist movement itself. To avoid such redundant spatiotemporally distant correlations, we propose dividing 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 it from whole body human poses. While the location and orientation of the human body can be trivial for most action classes, this is not the case for the hands. The hands do not conform to any particular 6D pose, and more significantly, they frequently and unpredictably alter their 6D pose over time. A comprehensive analysis of this phenomenon is provided in our Supplementary. In this regard, 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 single RGB frames 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 $P_t = \{P_t^{\text{left}}, P_t^{\text{right}}\} \in \mathbb{R}^{2 \times J \times 3}$ 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 framewise RGB features and $\mathcal{T}_r = \left\lfloor \frac{\mathcal{T}-1}{\Delta f} \right\rfloor \times \Delta f$.

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

To encode dense hand-pose sequences within micro-actions, we devise a trajectory-based pose encoder, 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 [34], whose

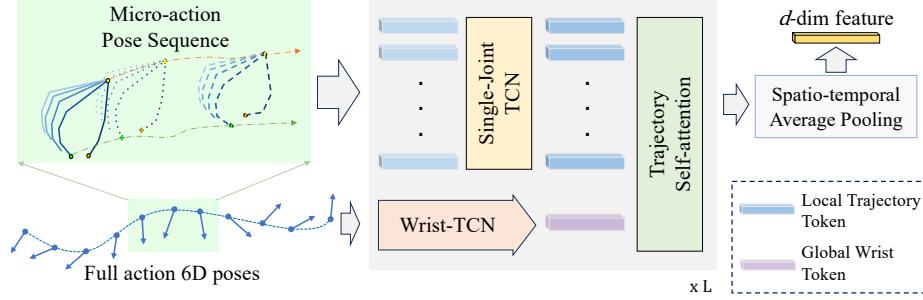


Fig. 4: Our Trajectory Encoder T , which operates on micro-actions, derives tokens with trajectory-based features and performs self-attention to encode the pose sequence into a feature vector. The Single-Joint TCN is a Temporal Convolutional Network [34] that processes the trajectories of all the joints 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.

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 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

In this section, we discuss incorporating sparsely sampled RGB frames into our HandFormer to better capture the scene semantics. A single frame is sampled from each micro-action, followed by generating an extended crop ($1.25\times$) around the hands. While the full image provides the 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 [54] or by using the corresponding hand pose projections.

Our multimodal tokenizer receives the frame feature and trajectory encoding of a micro-action as input Fig. 3, and performs multi-modal feature interaction to enhance their features for better RGB and pose tokens. This involves concatenating the frame features and the pose trajectories, projecting them into a shared PoseRGB feature space using an MLP, and then splitting the PoseRGB feature into two parts, which are added to the original RGB and Pose features.

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 [60], where we assign the same position to two tokens coming 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 via 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 by HandFormer. The pose and RGB modalities, serving as inputs, provide complementary information regarding a scene, capturing motion and interacting objects, respectively. To effectively utilize this information, we employ explicit verb and object supervision (\mathcal{L}_{verb} , \mathcal{L}_{obj}) through two additional learnable class tokens. Given that pose strongly correlates with verb class, the verb class token selectively attends to pose encodings. Similarly, as objects are reliably identifiable from RGB frames alone, the object class token exclusively attends 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 [24, 62], 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 with multiple views, Assembly101 [51] and H2O [33]. **Assembly101** [51] is a large-scale multiview dataset that features videos of procedural activities for assembling and disassembling 101 take-apart toy vehicles. This dataset has 1380 fine-grained actions resulting from a combination of 24 verbs and 90 objects. Additionally, 3D hand pose estimations are provided. The dataset consists of 12 temporally synchronized views — 8 static and 4 egocentric. The egocentric views exhibit low-resolution monochrome images, making it challenging even for human eyes to discern objects in view. As a result, we opt for fixed views in our RGB modality. **H2O** [33] features four participants performing 36 actions that involve 8 different objects and 11 verbs. 3D hand poses with 21 key points per hand are provided, along with 6D object poses, camera poses, object meshes, and scene point clouds. It consists of 5 temporally synchronized RGB views — 4 fixed and 1 egocentric. In our work, we use hand poses and RGB frames from the egocentric view only.

Implementation Details. Similar to [66], our pose input consists of 120 frames by setting $T' = 120$. 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 versions. Unless otherwise specified, we utilize the pre-trained ViT-g/14 and ViT-L/14 models from DINoV2 [46] followed by a linear layer without any fine-tuning as the frame encoder F for Assembly101 [51] and H2O [33], respectively. We also evaluate a lightweight alternative by fine-tuning ResNet-50 [28] in the Supplementary. 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}. During training, the frame encoder F is kept frozen except for the final linear layer.

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 strike a balance between 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 6 joints per hand (5 fingertips and the wrist) termed HandFormer-B/6.

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 [43] and an attention-based method ISTA-Net [66] — the two best

| Method | Pose | RGB | Assembly101 | | | H2O Action |
|------------------------|------|-----|--------------|--------------|--------------|---------------|
| | | | Action | Verb | Object | |
| MS-G3D [43] | ✓ | ✗ | 28.78 | 63.46 | 37.26 | 50.83 |
| ISTA-Net [66] | ✓ | ✗ | 28.14 | 62.70 | 36.77 | 89.09 |
| SlowFast [23] | ✗ | ✓ | - | - | - | 77.69 |
| TSM [39] | ✗ | ✓ | 35.27 | 58.27 | 47.45 | - |
| Cho <i>et al.</i> [13] | ✗ | ✓ | - | - | - | 90.90 |
| RGBPoseConv3D [20] | ✓ | ✓ | 33.61 | 61.99 | 42.90 | 83.47 |
| MS-G3D + TSM | ✓ | ✓ | 39.74 | 65.12 | 51.12 | - |
| | ✓ | ✗ | 28.80 | 65.33 | 36.28 | 57.44 |
| HandFormer-B/21 | ✗ | ✓ | 32.07 | 55.61 | 44.89 | 84.71 |
| | ✓ | ✓ | 41.06 | 69.23 | 51.17 | 93.39 |

Table 1: Quantitative comparison with state-of-the-art methods on Assembly101 [51] and H2O [33]. For H2O, 6D object pose is used by ISTA-Net during training and inference and by Cho *et al.* during training. ‘MS-G3D + TSM’ denotes a late fusion of the corresponding unimodal architectures.

performing skeleton-based methods on Assembly101 [51], as reported by [66]. We also employ video baselines TSM [39] and SlowFast [23], which emphasize efficiency and high temporal resolution, respectively. Additionally, we include Cho *et al.* [13], the state-of-the-art for the H2O dataset. For Assembly101, we replicate the results of the aforementioned methods 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], state-of-the-art in multi-modal action recognition with skeleton and RGB data.

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 in terms of action accuracy, which can be attributed to their high object recognition performance. 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 a 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

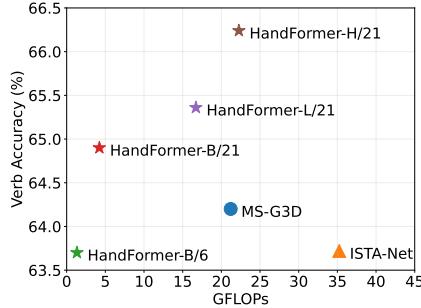


Fig. 5: Comparison of skeleton-based methods for verb recognition on Assembly101 [51]. Our method achieves state-of-the-art performance while utilizing significantly fewer FLOPs.

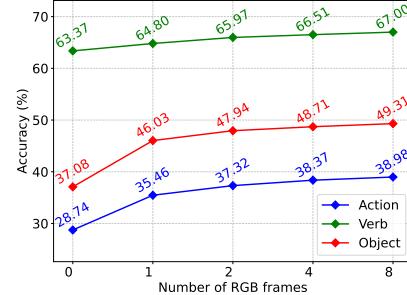


Fig. 6: Ablating number of RGB frames in HandFormer on Assembly101 [51]. With more RGB frames, verb recognition shows marginal gain, whereas object recognition shows improvement with diminishing returns.

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, while $J = 6$ variant is extremely efficient with comparable accuracy. HandFormer-H/21 is a combined variant of HandFormer-B and HandFormer-L with $J = 21$, which is 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 beneficial for object recognition and do 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]. 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 Are 3D hand poses an efficient alternative for multi-view?

Multi-view action recognition, while benefitting from precise hand-movement information in 3D space, processing all views with video models is expensive

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

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 performance (‘Single View + RGB’) matches the fusion of the two most informative views — *view 1* and *view 4*. Notably, our verb recognition accuracy is similar to the combination of all 8 RGB views while using $3\times$ fewer FLOPs (see Supplementary). Additionally, using hand pose in combination with an RGB view (‘Single View + RGB’) 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 also 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*. As a baseline video model, we train TSM [39] on both RGB views separately. As our method includes a 3D hand pose, which is obtained from 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 factors such as camera motion and the occlusion of interacting objects by hand. 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 compared to fixed RGB views, as demonstrated in [51]. We address this challenging scenario with our proposed multi-modal architecture and achieve state-of-the-art performance

| Method | Trained on | Tested on | Assembly101 [51] | | |
|------------|------------------|------------------|------------------|-------|--------|
| | | | Action | Verb | Object |
| TSM [39] | <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 |

Table 3: Cross-view performance of HandFormer 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 in Assembly101 and, therefore, are included in the video models.

| Method | Action | Verb | Object |
|---------------------------------|--------|-------|--------|
| TSM egocentric (fuse 4 views) | 33.80 | 59.00 | 46.50 |
| Egocentric (e_3) + Pose | 36.07 | 65.52 | 45.82 |
| Egocentric (e_4) + Pose | 35.56 | 65.79 | 45.20 |
| Egocentric (e_3+e_4) + Pose | 38.05 | 66.32 | 47.86 |

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

in egocentric action recognition on Assembly101. For this experiment, we used frame-wise TSM features provided by [51]. As depicted in Tab. 4, our model, using a single egocentric view (e_3 or e_4) outperforms the fusion of four egocentric views as reported in [51]. Moreover, fusing e_3 and e_4 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 impacts 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, and to expand to 11 joints, we additionally incorporate all the joints along the index finger and thumb, which are least affected by self-occlusion. We also show the effect of including 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 set number of pose frames, enlarging the window size for a micro-action will lead to a decrease in the number of micro-actions to aggregate, and conversely. In Tab. 6, we vary the micro-action length for verb recognition in the Assembly101 dataset [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.

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

Table 5: Keypoint 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 6: Micro-action length ablation for verb recognition in Assembly101 [51].

| Temp. Agg. | TCN | LSTM | Transformer |
|------------------|-------|-------|-------------|
| Verb Accuracy(%) | 62.95 | 63.34 | 63.70 |

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

| Multimodal Tokenizer | Reconstruction Loss | Verb & Object Loss | Action Accuracy (%) | |
|-------------------------|------------------------|-----------------------|---------------------|-------|
| | | | Assembly101 | H2O |
| ✗ | ✗ | ✗ | 38.98 | 85.95 |
| ✓ | ✗ | ✗ | 40.19 | 88.84 |
| ✓ | ✓ | ✗ | 40.24 | 89.26 |
| ✗ | ✗ | ✓ | 40.56 | 90.50 |
| ✓ | ✓ | ✓ | 41.06 | 93.39 |

Table 8: Ablating tokenization and different losses for action recognition on Assembly101 [51] and H2O [33].

Temporal Aggregation. After extracting features for micro-actions, 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.* fingertips and 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 7.45% for H2O [33].

6 Conclusion

With the growing interest in AR/VR and wearables, hand pose estimation has rapidly advanced, leading to the development of dedicated hardware independent of vision. As such, hand poses hold promise as a compact, domain-independent, rich modality to complement computer vision in the near future. To address the under-explored domain of using 3D hand poses as a modality for hand-object interaction recognition, we highlight the fundamental differences between hand poses and full-body human skeletons. We introduce HandFormer, a novel multi-modal 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 [26, 27], can encounter out-of-view scenarios and produce noisy poses. Although our experiments show that such estimated poses can still achieve good accuracy, further research can be done to explicitly address this phenomenon. We also assume that uniform sampling of RGB frames from each micro-action should provide us with good representations for understanding the semantic context, which is not true as all frames are not equally important in action. In such a case, adaptive frame sampling methods can be employed which we leave for future work.

Appendix

A Statistical Analysis: Full-body Skeleton vs. Hand

This section serves as an extension of Sec. 3 to statistically analyze the difference between hand poses and full-body poses. Action recognition datasets [40, 53], which primarily focus on full-body actions, often include actions involving partial body movements. These actions exhibit limited motion when viewed with respect to the entire skeleton, resulting in one or more relatively static joints. The change in the locations of moving joints with respect to such static joints can provide useful action cues. However, hand motions typically feature no such static reference points, as all the hand joints move together most of the time, making small changes in articulation to perform an action. To illustrate this difference, we randomly sample 1000 pose sequences from NTU RGB+D 120 [40] (full-body) and Assembly101 [51] (hands). We take the distance covered between two consecutive frames for each joint to form a distance array for the

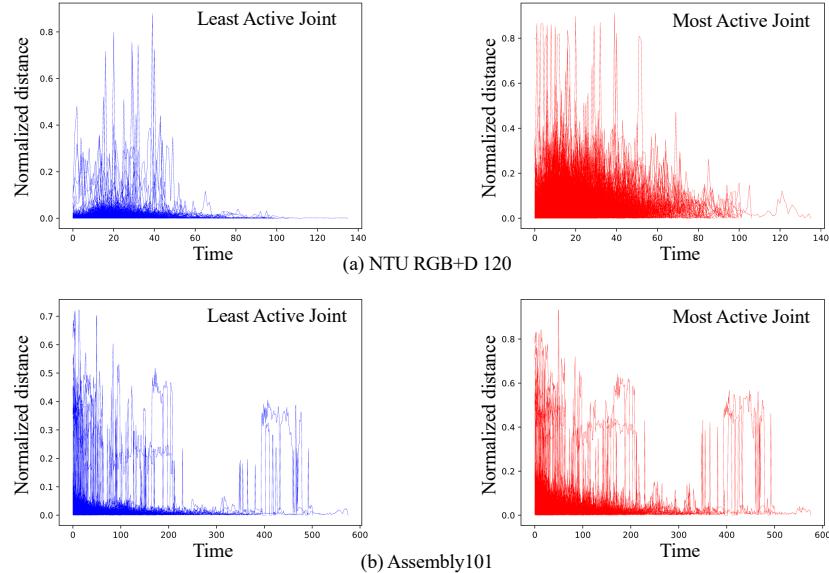


Fig. 7: With a random pool of 1000 sequences, we observe that the least active joints can be viewed as static reference points, showing minimal movement in NTU RGB+D 120. In contrast, Assembly101 exhibits subtler distinctions between the most active and the least active joints. The Pearson correlation coefficient (r) between the distance values for these two joints yields a high value (0.93) for Assembly101, while $r = 0.33$ for NTU RGB+D 120. These results suggest strong coupling among hand joints during motion, emphasizing the dominance of full-skeleton motion in hand poses. Our method leverages this understanding, balancing long-term motion patterns and short-term articulation changes by factorization.

corresponding joint j in a given sequence, which is determined by:

$$d_j(t) = \|P_j(t) - P_j(t-1)\| \quad (5)$$

Here, $d_j(t)$ is the distance covered by joint j at frame t in reference to the previous frame, $P_j(t)$ and $P_j(t-1)$ are the 3D pose coordinates for joint j at time t and $t-1$, respectively. $\|\cdot\|$ represents the Euclidean distance. Based on the sum of distances D_j for each joint j , we define the **least active joint** (static) and the **most active joint** (dynamic) for a particular sequence using the following equations:

$$\begin{aligned} D_j &= \sum_{t=1}^T d_j(t) \\ j_{sta} &= \arg \min_j D_j & j_{dyn} &= \arg \max_j D_j \end{aligned} \quad (6)$$

For all the sampled 1000 sequences, we take the temporal sequences $\{d_{j_{sta}}(t)\}_{t=1}^T$ and $\{d_{j_{dyn}}(t)\}_{t=1}^T$, normalize the distance values using the diameter of the corresponding skeleton, and plot the sequences separately in Fig. 7. As can be observed, compared to the distances covered by the most active joints in NTU RGB+D 120, the least active joints show significantly lower movement, effectively serving as the static reference points. On the other hand, the distinction in distance arrays between the most and the least active joints in Assembly101 is less pronounced. In addition, we calculated the Pearson correlation coefficient, denoted as r , between $\{d_{j_{sta}}(t)\}_{t=1}^T$ and $\{d_{j_{dyn}}(t)\}_{t=1}^T$ for all Assembly101 sequences, resulting in a value of **0.93**. Conversely, for NTU RGB+D 120, the corresponding correlation coefficient is **0.33**. This suggests strong coupling among hand joints during motion, and full-skeleton movement is more dominant in hand poses compared to full-body poses. Consequently, modeling dependencies between spatiotemporally distant joints is less effective for the highly dynamic hand motion (also discussed in Sec. 3). Therefore, by considering both long-term motion patterns and short-term articulation changes, our method facilitates efficient spatiotemporal factorization through micro-actions (Sec. 4.1). We also incorporate the full-skeletal motion from the entire action during micro-action encoding, using a global wrist token as a reference (Sec. 4.2).

B 2D vs. 3D Pose for Hand Actions

For skeleton-based action recognition, PoseConv3D [20] proposes using 2D poses as input, arguing that the quality of pose estimation is superior in 2D. By constructing 3D heatmap volumes from 2D poses and employing a simple 3D-CNN, they surpass state-of-the-art GCN-based methods that rely on 3D poses. Incorporating CNN-based modeling for the pose stream also facilitates seamless integration with the RGB modality. In this section, we assess this proposition specifically within the context of hand skeletons.



Fig. 8: Heatmaps for joints and limbs for (a) full-body poses and (b) hand poses.

Fig. 8 illustrates sample heatmaps from NTU RGB+D [40] and Assembly101 [51]. Keypoints in full-body human poses are often prominently situated, with minimal self-occlusion, and the subject is typically centered within the frame. As viewed in Fig. 8a, reducing pose dimensions to 2D does not significantly compromise detail; rather, it enhances input reliability by simplifying the pose estimation problem. However, this advantage diminishes when applied to hand poses. Hand poses present unique challenges, such as frequent self-occlusion and closer proximity of the keypoints, which are exacerbated by reducing the dimension to 2D. To empirically analyze this phenomenon, we evaluate HandFormer-B/6 with 2D and 3D poses for recognizing verbs in Assem-

| Method | Input Pose | Verb Accuracy (%) |
|-----------------|------------|-------------------|
| PoseConv3D [20] | 2D | 46.71 |
| HandFormer-B/6 | 2D | 58.92 |
| HandFormer-B/6 | 3D | 63.70 |

Table 9: Impact of using 2D vs. 3D poses as input for skeleton-based action recognition in hands. Experiments are done for verb recognition on Assembly101 [51].

bly101 [51] and report in Tab. 9. This analysis reveals about 5% difference in favor of the 3D input. Furthermore, PoseConv3D [20] introduces a CNN-based approach with 2D keypoints, which directly utilizes heatmaps from the pose estimator or generates Gaussian heatmaps from the 2D coordinates. However, feeding heatmaps to the model can diminish the clarity of keypoints, particularly when they are in close proximity, as is often the case with hand poses. Hence, PoseConv3D [20] performs poorly in recognizing hand actions, as evident in Tab. 9.

In summary, although skeleton-based methods represent a broader field for action recognition with poses, they often lack the necessary adaptation for directly addressing hand-specific actions. This demands dedicated research on hand poses for hand-object interaction understanding.

C Efficient Alternative for Frame Encoder

Except for our experiments on Assembly101 [51] with monochrome egocentric videos (Sec. 5.5), all our multimodal results are obtained using DINOv2 [46] features extracted from the input RGB frames. This approach enables us to assess the effectiveness of image-based foundation models in videos, leveraging all-purpose features from RGB frames, also achieving strong performance in cross-view generalization (Sec. 5.4). While using such frozen encoders for feature extraction can expedite training and mitigate the need for domain-specific modeling, the inference can be compute-heavy due to the large ViT backbones of foundation models. To alleviate this, we propose a pretraining scheme that allows

| Method Variant | Frame Encoder | Action | Verb | Object |
|----------------|---------------|--------|-------|--------|
| RGB-only | ViT-g/14 | 32.07 | 55.61 | 44.89 |
| | ResNet50 | 35.09 | 56.59 | 48.54 |
| Pose+RGB | ViT-g/14 | 41.06 | 69.23 | 51.17 |
| | ResNet50 | 41.99 | 69.28 | 51.96 |

Table 10: Comparison of different frame encoder options on Assembly101 [51]. Frame-wise TSM features from pretrained ResNet50 perform better compared to all-purpose features generated by DINOv2 with a ViT-g/14 backbone. RGB-only variant greatly benefits from the pretraining as it works with domain-specific features for action recognition. However, incorporating complementary pose modality reduces the gain from pretraining.

us to use a ResNet50 [28] backbone to extract image features, thereby reducing inference time. Specifically, we first train a TSM [39] model with a ResNet50 [28] backbone for action recognition, utilizing all action clips and then dropping the classification layer. This ResNet50 becomes the frozen image encoder in our proposed architecture, replacing ViT. During the training and inference of HandFormer, this TSM backbone operates as a true image model (ResNet50), as we employ it on individual frames without any channel shifting. The TSM features provided in the Assembly101 [51] are generated in this way, and we utilize them in our egocentric action recognition experiments (Sec. 5.5).

In Tab. 10, we present a comparison of the two backbone options for our frame encoder – ResNet50 from TSM and ViT-g/14 from DINOv2. The ResNet50 outputs, enhanced through pretraining within TSM, incorporate domain-specific features and temporal encoding via channel shifting during training. As a result, the RGB-only variant achieves a 3% higher action accuracy compared to using DINOv2 features alone. However, when introducing pose information, the image-based features are complemented by motion features, reducing the impact of motion understanding facilitated by the temporal shift mechanism of TSM in the ResNet50 encoder. Therefore, integrating pose data diminishes the pretraining advantage of ResNet50, resulting in a performance gap of less than 1%.

D Maintaining High Temporal Resolution at Low Cost

Our method is designed to perform action recognition efficiently in hand-object interaction videos. Obtaining efficiency in such a setup is challenging as we need to maintain a high temporal resolution to understand fine-grained hand movements that constitute the action. Therefore, we propose HandFormer using densely sampled pose frames and sparse RGB frames. In this section, we quantify the efficiency of this method compared to an alternative video model. As mentioned, understanding fine-grained hand motion demands a high temporal resolution to differentiate verb classes. For instance, relying on sparsely sampled frames may make actions like “*screwing*” and “*unscrewing*” indistinguishable. However, adopting a high temporal resolution with video models operating on

| Method | Component | GFLOPs | Count | Total GFLOPs |
|-----------------|----------------------|--------|-------|--------------|
| TSM [39] | - | - | - | 669.79 |
| | Pose Estimator [26] | 0.30 | 162 | |
| | Frame Encoder | 4.12 | 8 | |
| HandFormer-B/21 | Trajectory Encoder | 0.29 | 8 | 84.01 |
| | Multimodal Tokenizer | 0.01 | 8 | |
| | Temporal Transformer | 0.05 | 1 | |

Table 11: Comparison of FLOPs between Handformer and TSM [39] when both maintain a high temporal resolution at 60 fps. The number of frames is determined by the average action duration in Assembly101 [51], and we use eight non-overlapping micro-actions in our model.

RGB frames is challenging, primarily due to (*i*) the excessive computation associated with performing spatiotemporal operations on numerous frames, and (*ii*) the need to address redundancy in RGB frames to extract meaningful information.

In Tab. 11, we compare the FLOPs of our model vs. an efficient video model, TSM [39] with a ResNet50 backbone when both maintain a high temporal resolution. The results reveal that our model operates at about $8\times$ fewer FLOPs. As TSM has a 2D backbone and no 3D convolutions, it is expected to represent the lower bound for the computational cost of a video model at that temporal resolution. For our frame encoder, we opt for the efficient alternative as described in Sec. C. The average duration of fine-grained actions in Assembly101 [51] is 1.7 seconds. Following [51], we include an additional 0.5 seconds of context on either side, resulting in an average of $2.7 \times 60 = 162$ frames per action clip. We use $K = 8$ non-overlapping micro-actions, thus sampling 8 RGB frames and using the trajectory encoder eight times.

E Additional Details for Multimodal Training

Our training recipe for the multimodal HandFormer involves initializing the trajectory encoder with pretrained weights and utilizing hand-object ROI crop within the frame encoder — ensuring better use of pose and RGB, respectively.

E.1 Pretraining Trajectory Encoder

Encoding micro-action involves extracting RGB and pose features using frame encoder F and trajectory encoder T , respectively. While the frame encoder stays frozen and provides the appearance features, the trajectory encoder is learned and is expected to capture the hand motion. To effectively guide the trajectory encoder in achieving the desired encoding, we pretrain it for verb recognition solely using pose input. This approach leverages the inherent ability of pose data to capture hand motion, a key determinant of the verb while remaining agnostic to explicit information about interacting objects. This pretraining scheme leads to a better initialization of the trajectory encoder in multimodal HandFormer

| Frame Encoder | Trajectory Encoder Pretraining | Accuracy(%) | | |
|---------------|-----------------------------------|-------------|-------|--------|
| | | Action | Verb | Object |
| ViT-g/14 | ✗ | 39.79 | 67.40 | 50.69 |
| | ✓ | 41.06 | 69.23 | 51.17 |
| ResNet50 | ✗ | 40.47 | 66.00 | 51.10 |
| | ✓ | 41.99 | 69.28 | 51.96 |

Table 12: Initializing the trajectory encoder T with pretrained weights improves the overall performance with better verb recognition capability. Results are on Assembly101 [51] dataset. The initial weights for T are obtained by training the model to predict the verb classes from pose-only input.

for action recognition. In Tab. 12, we observe that initializing the trajectory encoder with pretrained weights leads to improved action recognition performance, particularly enhancing the recognition of verb classes.

E.2 Hand-Object Interaction Crop

In hand-object interaction (HOI) videos, the region of interest typically centers around the hands, capturing crucial information about the interacting object and the type of interaction. Leveraging 3D hand poses obtained through a readily available pose estimator [26], we project these poses onto RGB frames, extract the enclosing rectangle of the projected 2D pose, and expand it by 25% to define the ROI crop. However, relying solely on the cropped region can occasionally mislead the model for three potential reasons: *i*) failure of the pose estimator on certain frames, leading to the absence of useful features from the RGB frames, *ii*) the full object might not be visible when the crop is taken based on hand poses only, and *iii*) hand crops have limitations in capturing global changes compared to the full frames. Hence, to capitalize on both the localized interaction information of hand crops and the global contextual information provided by full frames, our model combines them both. If a valid hand crop is found, we take the full and cropped RGB frames, pass them through the frame encoder, average their features, and re-normalize them to unit norm. This full vs. HOI crop ablation is shown in Tab. 13, in which combining both perform better than the alternatives.

| Full Frame | HOI Crop | Accuracy(%) | | |
|------------|----------|-------------|-------|--------|
| | | Action | Verb | Object |
| ✓ | ✗ | 38.73 | 68.31 | 48.77 |
| ✗ | ✓ | 38.44 | 68.95 | 48.20 |
| ✓ | ✓ | 41.06 | 69.23 | 51.17 |

Table 13: Ablation study comparing full vs. HOI cropped RGB frames on Assembly101 [51]. Incorporating both full and cropped RGB frames allows for leveraging localized interaction details from hand crops and global contextual information from full frames, resulting in improved accuracy. HandFormer-B/21 is used with eight non-overlapping micro-actions.

F Efficiency Comparison with Shift-GCN

While MS-G3D [43] and ISTA-Net [66] show state-of-the-art performance for action recognition with hand poses, they are not efficiency-focused. Our HandFormer outperforms them with significantly fewer FLOPs. However, HandFormer-B/6 prioritizes efficiency while slightly trading off accuracy. Therefore, we implement and test an efficiency-focused baseline, ShiftGCN [11], for verb recognition on Assembly101 [51] and compare it to HandFormer-B/6 in Tab. 14. While Shift-GCN relies on graph shift operations and pointwise convolutions for efficiency,

| Method | GFLOPs | Verb Accuracy (%) |
|----------------|--------|-------------------|
| Shift-GCN [11] | 2.11 | 63.14 |
| HandFormer-B/6 | 1.33 | 63.70 |

Table 14: Comparison of HandFormer-B/6 with Shift-GCN, an efficiency-focused baseline for skeleton-based action recognition. Experiments are done for verb recognition on Assembly101 [51].

our model identifies the crucial joints, *i.e.* fingertips and wrist joint, and processes only these joints to reduce FLOPs substantially. As evident from Tab. 14, our model outperforms Shift-GCN while incurring lower FLOPs.

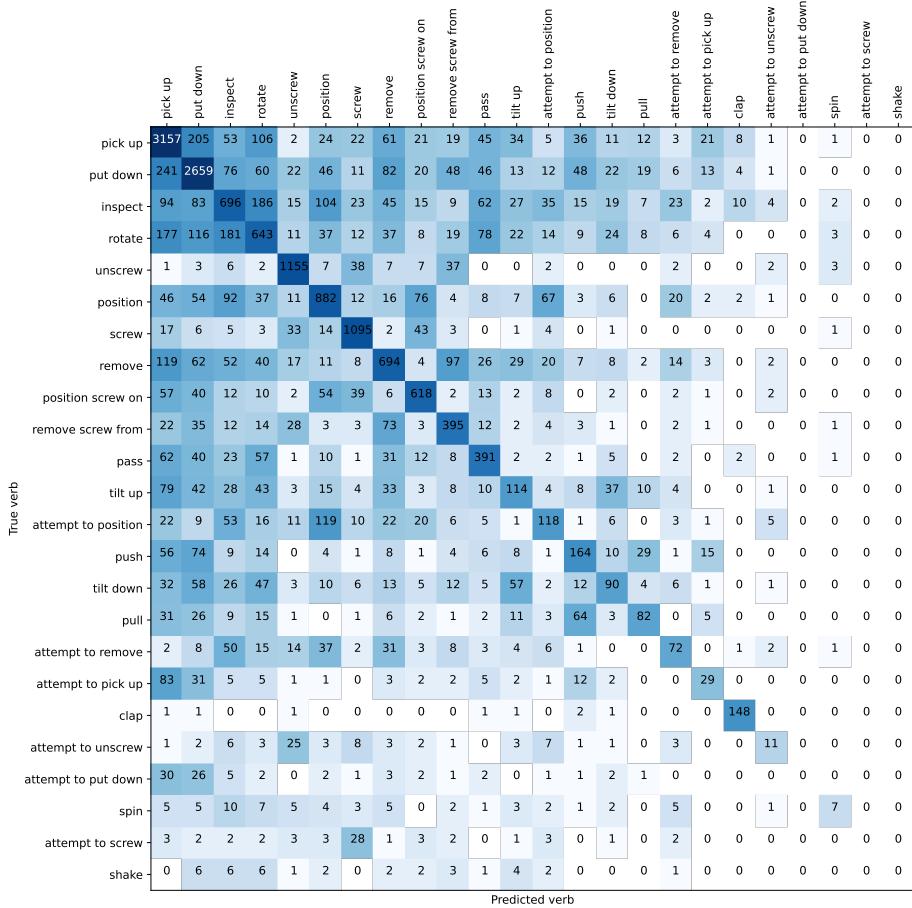


Fig. 9: Confusion matrix for pose-only verb recognition with HandFormer-L/21.

G Qualitative Analysis

In this section, we analyze the class-wise verb accuracy using the pose-only HandFormer, aiming to identify the model’s limitations. Furthermore, we examine the multimodal aspect of action recognition and its role in alleviating object misclassification.

G.1 Pose-only Performance

Fig. 9 displays the confusion matrix for verb classes using HandFormer-L/21 on the test set. Notably, *inspect*, *rotate*, *position*, and *remove* verbs present recognition challenges despite ample dataset samples. One potential explanation for this phenomenon is the shared presence of certain signature movements among these classes, which also occur in two head classes, namely, *pick up* and *put down*. Another interesting observation in the results is the frequent classification of ‘attempt to *x*’ classes as ‘*x*’. This is expected, as determining the successful completion of a task adds another layer of complexity to these classes, especially when relying solely on pose information without considering changes in the appearance of the interacting object throughout the clip.

G.2 Multimodal Fusion

To gain insights into how appearance information from RGB complements pose-based models in hand-object interaction scenarios, we analyze samples involving *put down* actions. In Tab. 15, we showcase the action classes predicted for these samples using our pose-only model, referred to as *Pose + 0 RGB*. In these samples, the model successfully detected the verb but struggled with object classification. This challenge arises due to similarities in articulations observed during tasks such as grasping a screwdriver and a screw or differentiating between a partially assembled toy and a completed one. These similarities lead to

| <i>Pose + 0 RGB</i> | Put down screwdriver | Put down screw | Put down partial toy |
|---------------------|---|---|---|
| RGB Sample |  |  |  |
| <i>Pose + 1 RGB</i> | Put down screw | Put down screwdriver | Put down finished toy |

Table 15: Action predictions by our model with and without sampling an RGB frame. Incorrect predictions are highlighted in red, while correct predictions are marked in green.

misclassifications by the pose-only model. However, introducing a single RGB frame, denoted as *Pose + 1 RGB*, enhances the model’s ability to correctly identify the relevant object by providing visual context. This observation highlights the limitations of recognizing actions, *i.e.* verb+object, solely from hand poses, emphasizing the importance of incorporating visual cues.

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) [3](#)
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](#)
3. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021) [3](#)
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) [3](#)
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) [3](#)
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) [3](#)
7. Caetano, C., Sena, J., Brémont, 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) [3](#)
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) [3](#)
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) [3](#)
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) [7](#)
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) [21, 22](#)
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) [3](#)
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) [4, 10](#)

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) [1](#)
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) [4](#)
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) [4](#)
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) [8](#)
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) [8](#)
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) [3](#)
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) [2](#), [3](#), [4](#), [10](#), [16](#), [18](#)
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) [3](#)
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) [3](#)
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) [1](#), [3](#), [10](#)
24. Girdhar, R., Grauman, K.: Anticipative video transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13505–13515 (2021) [8](#)
25. 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) [1](#)
26. 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) [1](#), [14](#), [19](#), [21](#)
27. 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) [1](#), [14](#)
28. 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) [9](#), [19](#)

29. 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) 3
30. 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) 3
31. 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) 3
32. 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) 3
33. 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) 1, 2, 3, 9, 10, 14
34. 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) 6, 7
35. 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) 1
36. 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) 3
37. 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) 4
38. 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) 3
39. 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) 3, 10, 12, 19, 20
40. 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) 2, 4, 15, 17
41. 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) 1
42. 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) 3
43. 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) 2, 3, 4, 9, 10, 21

44. 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) [1](#)
45. 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) [1](#), [2](#), [4](#)
46. Oquab, M., Dariseti, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) [9](#), [18](#)
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) [1](#), [3](#)
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) [3](#)
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) [2](#)
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) [3](#)
51. Sener, F., Chatterjee, D., Sheleporov, D., He, K., Singhania, 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) [1](#), [2](#), [3](#), [4](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [17](#), [18](#), [19](#), [20](#), [21](#), [22](#)
52. Sener, F., Singhania, D., Yao, A.: Temporal aggregate representations for long-range video understanding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. pp. 154–171. Springer (2020) [4](#)
53. 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) [2](#), [15](#)
54. 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) [7](#)
55. 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) [3](#)
56. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems **27** (2014) [3](#)
57. 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) [3](#)
58. 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) [3](#)

59. 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) [3](#)
60. 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) [8](#)
61. 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) [3](#)
62. Vondrick, C., Pirsiavash, 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) [8](#)
63. 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) [3](#)
64. 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) [4](#)
65. 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) [4](#)
66. 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) [2, 3, 4, 9, 10, 21](#)
67. 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) [1](#)
68. 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) [3](#)
69. 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) [2, 3](#)
70. 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) [3](#)
71. 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) [3](#)
72. 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) [3](#)