algorithms

MDPI

Article

A Unified Framework for Recognizing Dynamic Hand Actions and Estimating Hand Pose from First-Person RGB Videos

Jiayi Yang, Jiao Liang, Huimin Pan, Yuting Cai, Quanli Gao * and Xihan Wang *

School of Computer Science, Xi'an Polytechnic University, Xi'an 710600, China; 230721097@stu.xpu.edu.cn (J.Y.); 220721108@stu.xpu.edu.cn (J.L.); 230711005@stu.xpu.edu.cn (H.P.); 230711004@stu.xpu.edu.cn (Y.C.)
* Correspondence: gaoquanli@nwu.edu.cn (Q.G.); xihan_wang@xpu.edu.cn (X.W.)

Abstract

Recognizing hand actions and poses from first-person RGB videos is crucial for applications like human–computer interaction. However, the recognition accuracy is often affected by factors such as occlusion and blurring. In this study, we propose a unified framework for action recognition and hand pose estimation in first-person RGB videos. The framework consists of two main modules: the Hand Pose Estimation Module and the Action Recognition Module. In the Hand Pose Estimation Module, each video frame is fed into a multi-layer transformer encoder after passing through a feature extractor. The hand pose results and object categories for each frame are obtained through multi-layer perceptron prediction using a dual residual network structure. The above prediction results are concatenated with the feature information corresponding to each frame for subsequent action recognition tasks. In the Action Recognition Module, the feature vectors from each frame are aggregated by a multi-layer transformer encoder to capture the temporal information of the hand between video frames and obtain the motion trajectory. The final output is the category of hand movements in consecutive video frames. We conducted experiments on two publicly available datasets, FPHA and H2O, and the results show that our method achieves significant improvements on both datasets, with action recognition accuracies of 94.82% and 87.92%, respectively.

Keywords: gesture recognition; pose estimation; feature fusion

check for updates

Academic Editor: Laura Antonelli

Received: 21 April 2025
Revised: 21 June 2025
Accepted: 24 June 2025
Published: 27 June 2025

Citation: Yang, J.; Liang, J.; Pan, H.; Cai, Y.; Gao, Q.; Wang, X. A Unified Framework for Recognizing Dynamic Hand Actions and Estimating Hand Pose from First-Person RGB Videos. *Algorithms* **2025**, *18*, 393. https://doi.org/10.3390/a18070393

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Hand–object interaction tasks have a wide range of application areas, from virtual reality (VR) and augmented reality (AR) to human–computer interaction. Also, recognizing hand movements from first-person RGB videos has received significant attention from researchers due to the rich visual information provided by RGB videos. On the one hand, the inherent complexity of hand articulation, combined with frequent occlusions (either self-occlusions between fingers or occlusions by objects), severely impedes the capture of complete hand information during the feature extraction stage. On the other hand, compounding this difficulty, hand movements within complex RGB video scenes constitute inherently continuous temporal sequences. Effectively comprehending spatiotemporal action demands the precise capture of nuanced inter-frame action coherence and subtle motion dynamics, necessitating robust temporal modeling capabilities. Notably, there is a high correlation between 3D hand pose estimation and hand movement recognition [1], yet there are relatively few joint studies on the two. To address this issue, recent studies have started to explore frameworks that combine hand pose estimation with action recognition [2–5]

for more accurate action understanding. However, the task still faces many challenges. Tekin et al. [3] proposed an end-to-end deep learning model, H+O, which, for the first time, modeled hand pose estimation, object pose and interaction state understanding in a unified way, emphasizing that the interaction between the two is crucial for action understanding in a first-person perspective. Building upon dual-task frameworks, Yang et al. [5] proposed a multi-order multi-stream feature analysis method using weakly supervised network learning to learn gesture and multi-order motion information from the intermediate feature maps of the video, which is capable of learning gesture recognition and 3D hand pose estimation, even when only pose labels are available. Crucially, their paradigm capitalizes on the mutual reinforcement between hand pose semantics and action contexts, thereby partially resolving occlusion-induced ambiguities. In parallel, Kwon et al. [2] extended Tekin et al.'s [3] foundation via a topology-aware graph convolutional network (TA-GCN), explicitly modeling holistic hand–object interdependencies. By integrating gesture kinematics, object-state dynamics, and their bidirectional relational constraints, their approach significantly advances interaction modeling precision. Despite these methodological innovations, temporal coherence across sequential frames remains inadequately optimized, limiting contextual continuity in video analysis. Consequently, extracting accurate hand motion semantics from first-person RGB footage persists as a formidable challenge.

In this study, we present a framework for recognizing hand movements and hand gestures from first-person RGB videos. The framework consists of two main modules: the Hand Pose Estimation Module and the Action Recognition Module, as shown in Figure 1. In the Hand Pose Estimation Module, individual video frames are first passed through a feature extractor and then passed into a series of transformer encoders. We construct a dual residual network structure and use a multi-layer perceptron to predict the hand pose results and object categories for each frame. High-level semantic features and low-level spatial details are jointly modelled via two residual paths. The capture of fine-grained spatial information is enhanced using a spatial detail feature extraction component while retaining global feature consistency. The hand pose results, object categories, and feature information of each frame are concatenated for the subsequent action recognition task. In the Action Recognition Module, the vectors of each frame are aggregated by means of a multi-layer transformer encoder. The temporal information of the hand between video frames is captured, the motion trajectory is obtained, and, finally, the hand action categories in consecutive video frames are output. We evaluate our approach on the first-person hand motion benchmark datasets FPHA [1] and H2O [2]. We summarize the contributions as follows:

(1) We introduce a unified framework for action recognition and pose estimation, enabling prediction of each frame hand pose and object category information, and, at the same time, it is able to output the hand movement categories in consecutive video frames.

(2) We design two interconnected modules to capture the spatial and temporal information of the hand in the video, namely the Hand Pose Estimation Module (HPEM) and Action Recognition Module (ARM). On one hand, HPEM is used to capture the spatial position of the hand in a single-frame RGB image. The pose estimator through a double residual fusion structure is combined with a spatial detail feature extraction component. The key points of the hand in each image frame are further estimated to obtain spatial information. On the other hand, ARM is used to aggregate the features of consecutive video frames in the temporal dimension and obtain the complete motion trajectory of the hand and ultimately obtain the action classes by the action classifiers.

(3) Comprehensive evaluations of two first-person benchmarks (FPHA, H2O) demonstrate state-of-the-art performance.

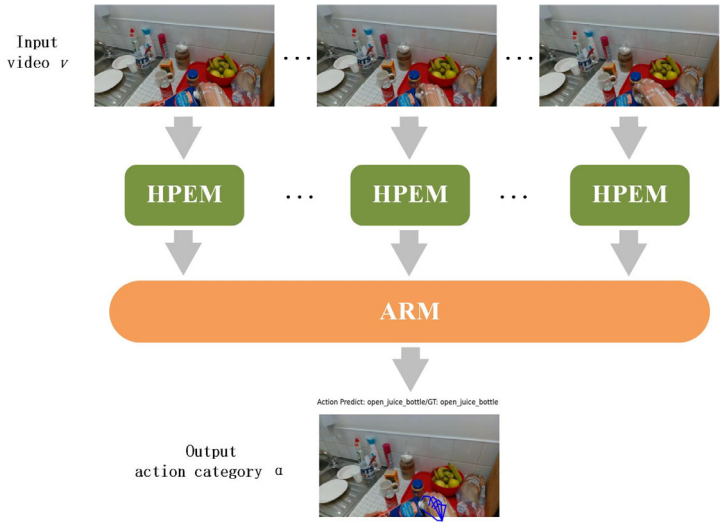


Figure 1. Schematic diagram of the framework process. The framework first processes each frame in the input first-person-view RGB video υ to estimate the hand pose for each frame via HPEM and determine the type of object involved within each frame. Subsequently, the processed frames are passed to the ARM, which aggregates the information from all the frames, captures the temporal dynamics of the hand movements, and ultimately outputs the recognition results of the hand movements.

2. Related Work

In this section, we present research related to our work: (1) estimation of 3D hand pose from RGB images/videos, (2) prediction of action recognition from RGB images/videos, (3) vision tasks in the transformer

2.1. 3D Hand Pose Estimation from RGB Image/Video

The ubiquity of RGB cameras has accelerated progress in estimating hand poses from RGB imagery. Hand pose estimation has emerged as a focal area in computer vision, driven by advancements in monocular RGB-based methodologies [6–15]. Zimmermann et al. [15] introduced a neural framework combining pre-trained 3D joint models with detected 2D key points to estimate hand poses from single RGB inputs. Lin et al. [9] proposed a novel architecture that combines a graph convolutional network with the transformer architecture to achieve simultaneous pose estimation and hand mesh reconstruction. To resolve depth uncertainty, Iqbal et al. [7] leveraged a 2.5D heatmap encoding strategy within a convolutional network, enhancing positional precision through multi-dimensional feature fusion. Moon et al. [11] employed a hierarchical framework utilizing pixel-wise heatmap predictions for 2D joint localization and depth regression. Spurr et al. [14] addressed the challenge of data scarcity in 3D hand pose estimation by proposing a weakly supervised framework. Robust pose estimation can be achieved through kinematic constraints without the need for full 3D annotation.

Related research methods have been extended to 3D hand pose estimation from continuous video data [1,16–18]. Temporal continuity in video streams offers enhanced contextual cues compared to static frames, capitalizing on motion patterns and kinematic coherence. Chen et al. [17] introduced a self-supervised framework for inferring 3D hand meshes and joint trajectories from unannotated videos while eliminating dependency on manual 3D supervision. Garcia-Hernando et al. [1] introduced a large-scale first-person RGB-D dataset that addresses a critical gap in first-person movement analysis. Depth sensors and inverse kinematics are also utilized to generate accurate 3D joint annotations, avoiding the reliance on error-prone manual annotations. Research on gesture recognition and fine-grained motion understanding is enhanced. Cai et al. [16] advanced temporal

modeling via graph convolutional architectures, leveraging spatiotemporal coherence in sequential 2D joint data to jointly estimate body and hand poses.

2.2. Action Recognition from RGB Image/Video

Action recognition constitutes a critical subfield of computer vision, focusing on the automated interpretation of human behaviors from images or videos. With the development of deep learning techniques, a large number of researchers have begun to explore fine-grained action recognition for specific body parts (e.g., hand, lower limbs) [19–23]. Since temporal information is usually required for hand action recognition, many advances have been made on architectures that emphasize temporal information processing [24–30], such as using various convolutional networks [30,31], using LSTM [27], and using dual-stream convolutional networks [24,26,28]. Carreira et al. [24] proposed a dual-stream convolutional architecture that extends 2D convolutional networks to 3D and incorporates optical flow information, verifying the complementarity of optical flow information with RGB data. Shi et al. [28] designed an adaptive graph convolution framework based on a temporal model. The framework adopts a dual-stream architecture and improves the accuracy of skeletal action recognition by modelling skeletal sequences hierarchically. Feichtenhofer et al. [25] introduced a multi-rate classification and detection of video actions framework, in which a low-frame-rate pathway captures spatial semantics while a high-frequency branch isolates motion dynamics, optimizing recognition efficiency. A study by Simonyan et al. [29], also built on dual-stream convolutional networks, proposes a dual-stream deep convolutional network framework containing both spatial and temporal networks, which greatly improves the accuracy of deep video classification.

2.3. Transformers in Vision

Transformer is a deep learning model originally proposed by Vaswani et al. [32] for sequence-to-sequence tasks. Due to its parallel processing capability and efficient self-attention mechanism, transformer excels in processing long sequences and capturing long-range dependencies, leading to major successes in natural language processing tasks. At the same time, its success has catalyzed cross-domain adaptations, particularly in visual computing, where it has demonstrated efficacy across diverse tasks including image detection and pose estimation. Carion et al. [33] proposed an end-to-end framework employing transformer for target detection tasks, effectively addressing conventional CNNs' limitations in modeling inter-object relationships across distant spatial regions. Subsequent innovations by Lin et al. [9,34] and Hampali et al. [35] further validated transformer's potential through convolution-free implementations for pose estimation, achieving state-of-the-art accuracy. Despite these advancements, due to quadratic complexity growth with input resolution, the transformer's quadratic complexity with respect to input size poses significant computational challenges.

3. Methodology

In this study, we propose a unified framework that predicts each frame hand key point information along with object category information from first-person RGB video, while also being able to output hand action categories across consecutive video frames. The framework is illustrated in Figure 2, comprising two modules: the Hand Pose Estimation Module (HPEM) and the Action Recognition Module (ARM). In the Hand Pose Estimation Module, each video frame is first passed through a feature extractor and then fed into multiple transformer encoders. We construct a dual residual network structure and use a multi-layer perceptron to predict the hand key point coordinates and object categories for each frame. The hand pose results, object categories, and feature information of each frame

are concatenated for the subsequent task. In the Action Recognition Module, the vectors of each frame are aggregated by a multi-layer transformer encoder. The temporal dynamics of the hand between video frames are captured to infer or estimate the motion trajectory, ultimately enabling the classification of hand action categories over the video sequence. In the remainder of this section, we examine each sub-module in more detail.

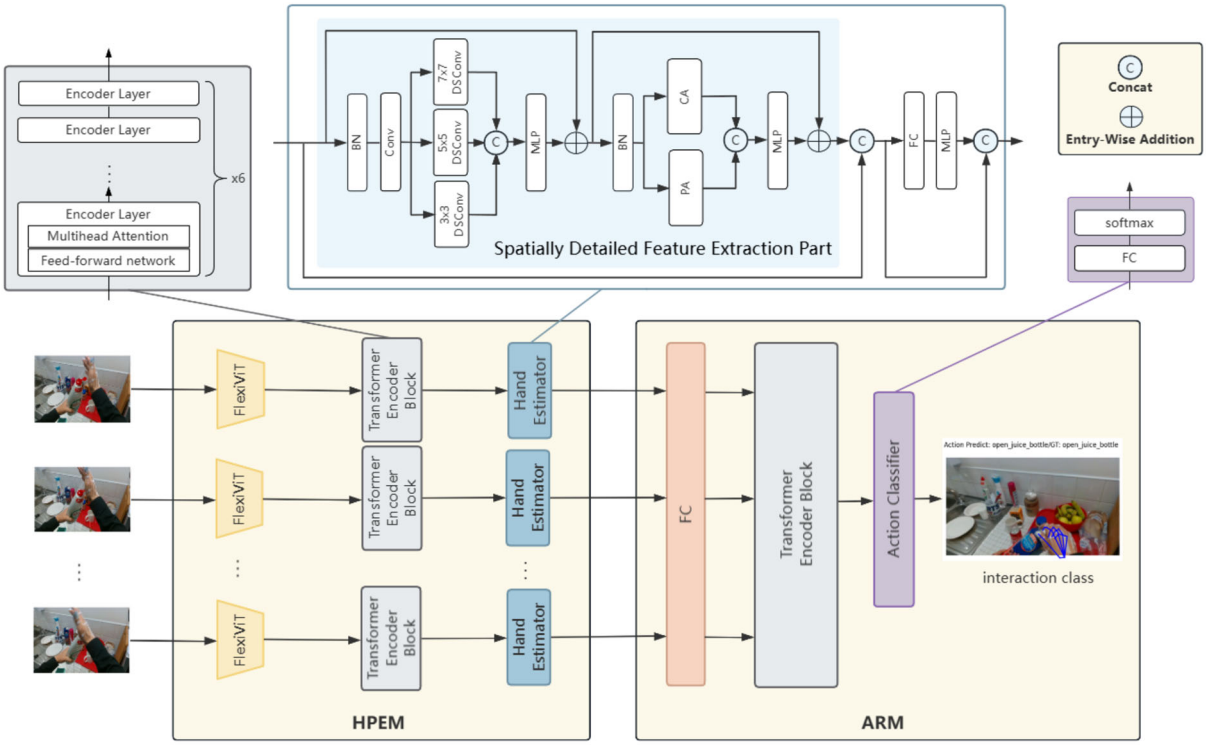


Figure 2. Schematic of the overall framework. Our framework comprises two main modules: the Hand Pose Estimation Module (HPEM) and the Action Recognition Module (ARM). For the input video υ, the spatial position of the hand in a single video frame is captured by HPEM. The spatial information is obtained by estimating the key points of the hand in each frame using a dual residual network structure and multi-layer perceptron. The ARM captures the temporal information of the hand between video frames by aggregating the information of each frame through a multi-layer Transformer encoder and ultimately outputs the class of hand actions in consecutive video frames.

3.1. Hand Pose Estimation Module

To accurately capture the spatial positioning of the hand in an RGB video, the task of predicting hand key point coordinates and object category information for each frame from a given first-person RGB video is implemented. We designed the lightweight Hand Pose Estimation Module based on transformer and dual residual network structure. The module is described in detail in this subsection.

Due to a large amount of redundant information between neighboring frames of the video, the direct computation cost of the transformer-based approach is high. If the whole video is directly fed into the transformer-based model for processing, it will not only bring high computational overhead but may also cause memory overflow problems when processing multiple frames at the same time. To mitigate this, we divide the original video sequence υ into j consecutive subsets, each of which set $= \{\text{set}_i\}_{i=1}^{j}$ is processed as an independent input to the network. This approach reduces the computational complexity while ensuring temporal continuity, effectively preserves the ability to perceive dynamic gesture changes, and, thus, improves the overall inference efficiency. Meanwhile, from the perspective of model training efficiency and performance, this approach helps to control the amount of data in each batch of inputs and achieve stable training and efficient inference

under limited hardware resources. For each input frame t ∈ set ⊂ $R^{W\times H\times 3}$, the pre-trained FlexiViT [36] architecture is used as our backbone network to extract feature information from each frame, where W represents the width and H represents the height of the input image. Specifically, using 16 × 16 image block segmentation, each image block is mapped to an embedding space of dimension 192 by linear projection, thus forming an embedding vector with semantic information for subsequent feature learning. A 12-layer transformer encoder structure is used, with each layer containing 3 parallel attention heads. While maintaining a low computational complexity, multi-layer image features are gradually extracted to enhance the feature representation capability.

In order to efficiently model the temporal dependence in the input video sequence, we generate the position information of the input sequence by means of fixed sine/cosine position encoding [32]. This enables the model to more effectively capture the relative positions of elements within the time-series data. Each frame of the visual feature sequence obtained through the FlexiViT backbone and its corresponding position encoding is passed to the transformer Encoder Block for processing to obtain the global context features $f_{glo}(t)$. The transformer Encoder Block is one of the core components of the network responsible for modelling long-range dependencies and mining contextual associations between features. Also, Layer Normalization is applied prior to the attention and feed-forward operations to stabilize the training process. A transformer Encoder Block is composed of six stacked encoder layers, each comprising a multi-head deformable self-attention layer with eight heads and a feed-forward network with a dimension of 2048. A transformer Encoder Block significantly enhances its ability to capture spatiotemporal dependencies and further enriches the feature representation.

In order to further strengthen the ability to perceive the local structure of the hand and the region around the key point, more accurate spatial details are extracted on the basis of preserving the original features, and the problems of gradient vanishing and gradient explosion in deep network training are effectively mitigated. We designed a hand pose estimator based on a dual residual network structure, including two parts: a Spatially Detailed Feature extraction part and an estimation part. For the Spatially Detailed Feature Extraction Part, multi-scale parallel convolution operation and enhanced parallel attention operation are used. The standard convolution is able to expand the sensory field and extract global features, and the multiscale parallel depth-separable convolution operation captures different levels of image information. Here, 3 × 3, 5 × 5, and 7 × 7 sized convolution kernels are used and then the results are fused. In addition, to improve the model's attention to the key features of the hand, a parallel structure is used here to fuse different types of attention mechanisms, including channel attention and pixel attention. Among them, the pixel attention focuses on the accurate extraction of hand position-related information, while the channel attention can efficiently extract global information while adjusting the feature channel dimensions. Specifically, we use BatchNorm for normalization to splice the results of these two attention mechanisms along the channel dimension. An MLP is applied to project the concatenated features back to the original input dimensionality.

We concatenate the Spatially Detailed Features $f_{det}(t)$ obtained from the above operation with the global context features $f_{glo}(t)$ by residual concatenation to obtain $f'(t)$

$$f'(t) = \Big\| f_{det}(t) + f_{glo}(t) \tag{1}$$

where ‖ is the concatenation operation. Finally, a multi-layer perceptron is used to accurately predict the coordinates of the hand key points in each frame from the features $f'(t)$:

$$H_t = \left(H^{pre}, H^{pre}_{dep}\right) = FC\ [\ MLP\ (f'(t))\ ] \tag{2}$$

where H_t is the concatenation of the H^{pre} and H^{pre}_{dep}, H^{pre} denotes the joint coordinates in the 2D image, H^{pre}_{dep} denotes the joint depth to the camera. The FC layer maps the features to the new space, and, similarly, we predict the object class from the feature $f'(t)$ and fuse the object category prediction as auxiliary information into the action recognition module.

Finally, inspired by [37], we concatenate the hand pose estimates, object category predictions, and frame-wise visual features $f'(t)$ to construct the input representation for the action recognition module. Figure 3 shows the importance of object category information for disambiguation. This fusion strategy preserves information integrity while maintaining computational efficiency, explicitly bridging high-level semantic cues with low-level visual details, thus improving discriminative capability for complex hand–object interactions. Specifically, for a frame $t \in set \subset R^{W \times H \times 3}$ with $set = \{set_i\}_{i=1}^{j}$, $F_{HPEM}(t)$ is the result of HPEM calculation:

$$F_{HPEM}(t) = \| H_t + O_t + f'(t) \tag{3}$$

where $\|$ is the concatenation operation; O_t is the probability distribution of object classification.

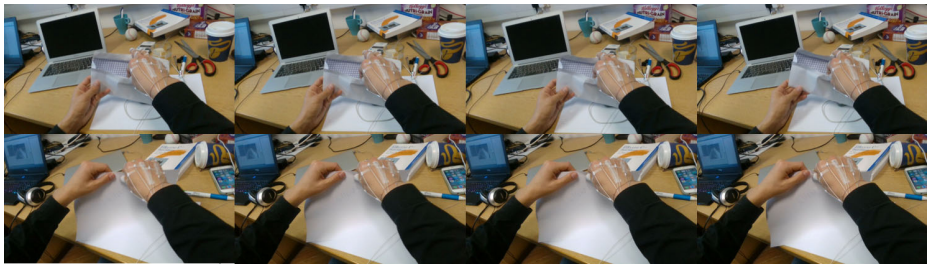


Figure 3. Partial sequences of the action classes 'take_letter_from_envelope' (**first row**) and 'tear_paper' (**second row**) are shown in the FPHA dataset. It can be seen that when the gestures of the hands are almost the same, the difference in the objects touched will lead to the difference in the final action categories. This shows that object category information is important for eliminating ambiguity and improving the accuracy of action recognition.

3.2. Action Recognition Module

To effectively capture the dynamic variations in hand movements in the time dimension, the task of action recognition from a given first-person RGB video is achieved. The Action Recognition Module was designed using the transformer architecture as its foundation. The input to this module is each frame $F_{HPEM}(t)$ obtained from HPEM. The vectors of each frame are aggregated by the transformer Encoder Block, and the final output is the class of hand movements in consecutive video frames. We will describe each step in detail in this subsection.

Before proceeding with the action recognition task, we first perform a linear transformation of the features to compress the high-dimensional features into a uniform low-dimensional space. Also, we follow [38,39] to introduce a trainable token, which fuses the global token with the linearly transformed features by aggregating the global information of the video υ for action classification. A transformer Encoder Block, composed of six stacked encoder layers, is used to aggregate the vectors for each frame. The temporal information of the hand between video frames is captured to obtain the complete motion trajectory of the hand. The final hand movement categories in consecutive video frames are output by the movement classifier. Specifically, we use the first token α_{out} of the output sequence of the transformer Encoder Block to classify and obtain the probability distribution over each action category, defined as follows:

$$A(v) = \{p(a_1), p(a_2), \ldots, p(a_n)\} = softmax(FC(\alpha_{out})) \tag{4}$$

where n represents the total count of the predefined categories of hand actions.

3.3. Loss Functions

Our total training loss consists of three components: (1) object label prediction loss, (2) hand pose estimation loss, and (3) action recognition loss.

For the object label prediction task, HPEM predicts the object labels for each frame, and we perform supervised learning for this task by minimizing the cross-entropy loss function L_{obj}. L_{obj} defines the difference between the actual object labeling distribution $o^{gt} = \left\{o_i^{gt}\right\}_{i=1}^{n_a}$ and the predicted object labeling probability distribution $o^{pre} = \left\{o_i^{pre}\right\}_{i=1}^{n_a}$.

$$L_{obj}(t) = -\sum_{i=1}^{n_a} O_i^{gt} log\left(O_i^{pre}\right) \tag{5}$$

For the hand pose estimation, we perform supervised learning with L1 loss, aiming to minimize the L1 loss during the training process. The loss function $L_{hand}(I)$ compares the difference between the predicted joint coordinates $\left(H^{pre}, H_{dep}^{pre}\right)$ and the true value $\left(H^{gt}, H_{dep}^{gt}\right)$. The loss function is articulated as follows:

$$L_{hand}(t) = \frac{1}{J}\left(\left\|H^{pre} - H^{gt}\right\|_1 + \alpha_1\left\|H_{dep}^{pre} - H_{dep}^{gt}\right\|_1\right) \tag{6}$$

where J represents the total number of joints in the hand. H^{pre} and H_{dep}^{pre} are the coordinates and depth of the joints in the 2D image plane predicted by the model, respectively. H^{gt} and H_{dep}^{gt} are the actual 2D joint coordinates and depth, respectively. α_1 is a parameter used to balance the different weights between 2D coordinate losses and depth losses.

For the action recognition task, ARM aggregates frame-wise feature representations to predict the video's action class, and supervised learning is applied to train this task by minimizing the cross-entropy loss function. L_{action} defines the difference between the actual action label distribution $a^{gt} = \left\{a_i^{gt}\right\}_{i=1}^{n_a}$ and the predicted action label probability distribution $a^{pre} = \left\{a_i^{pre}\right\}_{i=1}^{n_a}$:

$$L_{action}(v) = -\sum_{i=1}^{n_a} a_i^{gt} log\left(a_i^{pre}\right) \tag{7}$$

The total loss function of the total training loss combines the above three losses, using the weighting parameters α_2 and α_3 to control the relative importance of each part of the optimization process. We define the total training loss L_{total} as follows:

$$L_{total} = L_{action}(v) + \frac{1}{T}\sum_{\overline{v}=seg_t(v)}\sum_{t\in\overline{v}}\left(\alpha_2 L_{obj}(t) + \alpha_3 L_{hand}(t)\right) \tag{8}$$

where α_2 and α_3 are parameters that balance the object label prediction loss term and the hand pose estimation loss term. By employing the combined loss function, the model can jointly optimize object label prediction and hand pose estimation, thereby ensure the accuracy of the final action recognition task and further enhancing the overall stability and performance of the network.

4. Experiments

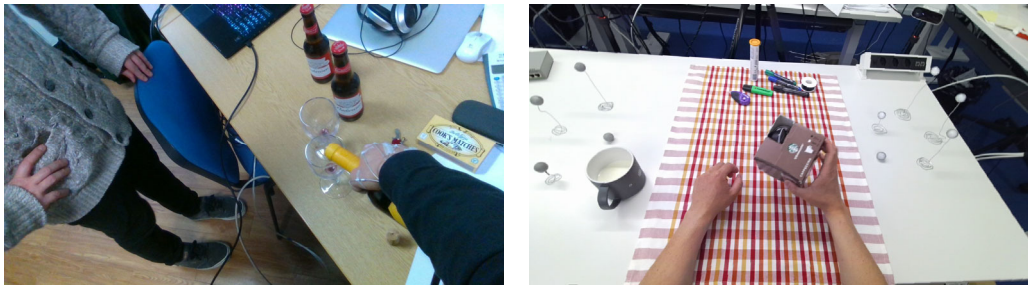
4.1. Experiment Details

We use PyTorch (version: 1.11.0) to implement our network. The FlexiViT network is initialized using pre-trained weights. We set α_1= 200, α_2 = 0.5, and α_3 = 1 to balance the loss terms for different tasks. All models in the experiment were trained for 40 epochs. The initial learning rate is set to 3×10^{-5} with a halving of the learning rate every 15 epochs and using the Adam optimizer [40] to train the framework. The entire network was trained and evaluated using an RTX 3090 GPU with a batch size of 1. The input image size was set to 270 $\times$ 480 while the input image was augmented with online data by randomly adjusting the hue and contrast and by adding random stochastic rotations, translations, and motion blur.

4.2. Datasets and Metrics

We conducted experiments and tests on two first-person hand public datasets, FPHA [1] and H2O [2].

FPHA: This dataset contains more than 100 K RGB-D frames annotated with 3D gestures. It features six subjects performing 45 action classes involving 26 distinct objects, as shown in Figure 4a. The FPHA dataset includes annotations for a single hand and object. When categorized by action, the training set comprises 600 videos, while the test set contains 575 videos.



(**a**) (**b**)

Figure 4. Dataset image presentation. (**a**) Image of FPHA dataset; (**b**) image of H2O dataset.

H2O: This dataset is an interaction scene between two hands and an object, containing 4 themes, recording 36 actions, and 8 different objects, providing unlabeled 3D annotations for both hands, as shown in Figure 4b. Although the H2O dataset provides multi-view images, we only utilize the images captured from the first-person viewpoint in our experiments. We adopt the methodology outlined in [2] for both training and evaluation, with the training set comprising 569 videos and the test set including 242 videos.

In our quantitative evaluation, we adopt task-specific performance metrics.

For the action recognition task, we computed the action recognition accuracy of the model on the test set as a primary measure of classification performance. Specifically, we compared the action categories predicted by the model on each video sample in the test set with the corresponding real labels and counted the ratio of the number of correctly recognized samples to the total number of samples, so as to evaluate the model's ability to discriminate different action categories. This metric can intuitively reflect the model's performance in semantic understanding, especially whether it can accurately capture the dynamic features and make correct recognition when facing complex hand–object interaction actions.

For the hand pose estimation task, we adopt Mean Endpoint Error (MEPE) as the main evaluation criterion. MEPE is defined as the average of the Euclidean distance between the

3D joint coordinates predicted by the model and the ground-truth joint coordinates. This metric can effectively quantify the accuracy of the model in the spatial dimension for the reproduction of hand structure and posture. The accuracy of hand key point localization is objectively assessed by calculating the MEPE values of all samples on the test set.

4.3. Experimental Results

4.3.1. Comparison with State-of-the-Art Hand Pose Estimation Methods

We compare our model with state-of-the-art hand pose estimation methods and evaluate it on the FPHA and H2O test sets. Table 1 presents a summary of the hand pose estimation results on the H2O dataset, demonstrating that our approach outperforms all existing methods. Figures 5 and 6 illustrate the qualitative results of our framework applied to the test sets from the FPHA and H2O datasets, respectively.

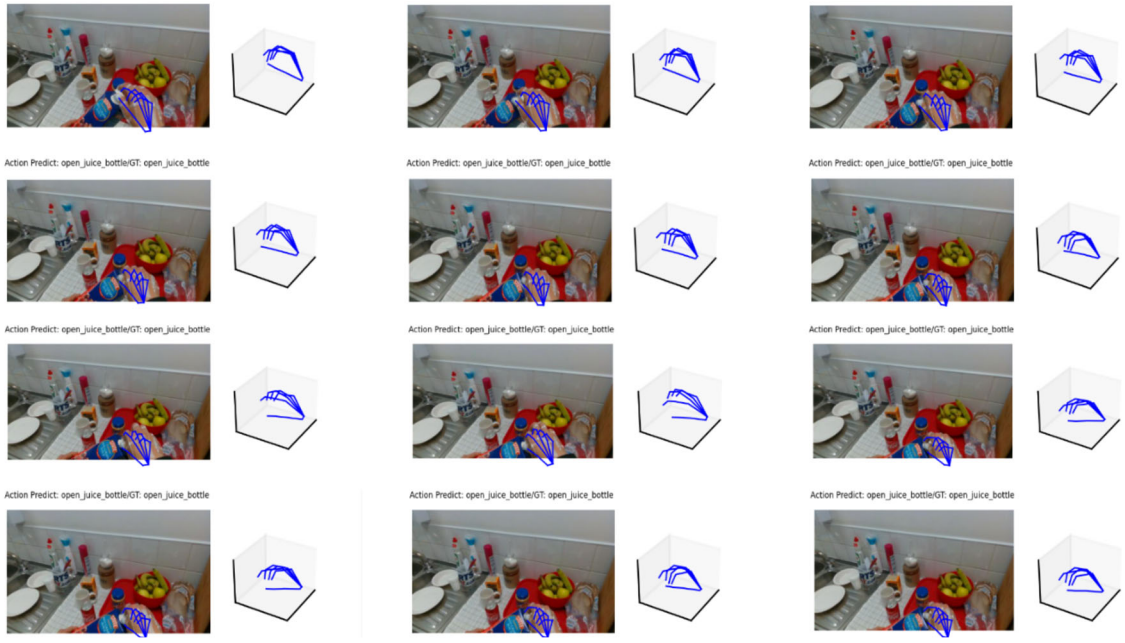


Figure 5. An example of hand posture predicted by our framework on the FPHA dataset.

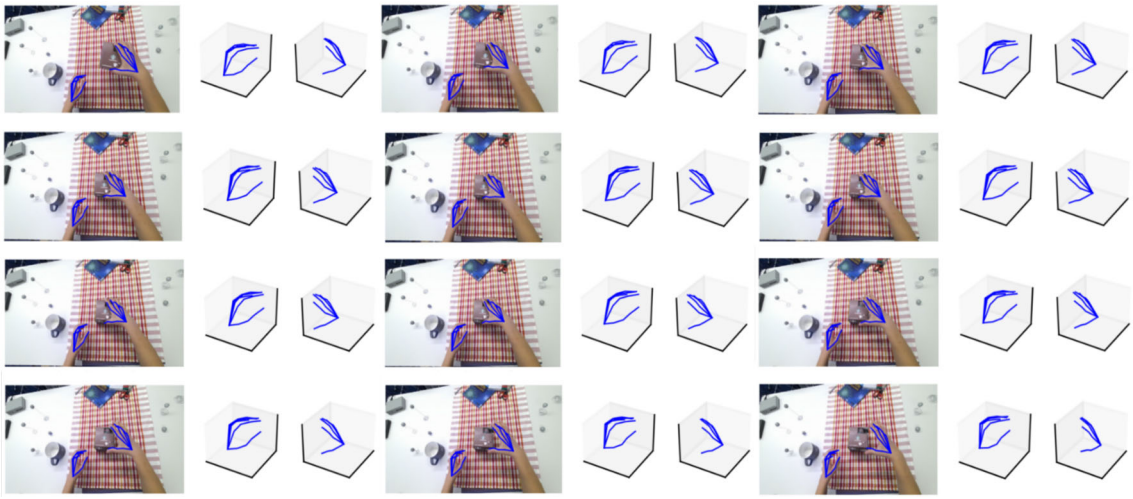


Figure 6. An example of hand posture predicted by our framework on the H2O dataset.

Table 1. The quantitative results of manual pose estimation in the H2O test set, MEPE, measured in millimeters.

Method	Left	Right
H+O [3]	41.42	38.86
LPC [41]	39.56	41.87
H2O [2]	41.45	37.21
HTT [4]	35.02	35.63
THOR-NET [42]	36.8	36.5
Ours	**31.16**	**35.21**

4.3.2. Comparison with State-of-the-Art Action Recognition Methods

On the FPHA dataset, we compared the proposed method with a number of cutting-edge methods. These methods are based on different network architectures and feature fusion strategies and have a high impact in the field of first-view hand movement recognition. Among the 17 baselines, Two-stream [26], Collaborative [5] and Joule-color [43] are based on CNNs. HAN-2S [44], ResGCNeXt [45], SBI-DHGR [46], K. Prasse et al. [47], GCN-BL [48] are based on the skeletons. FPHA [1] and H+O [3] are based on the RNN. Li et al. [49] is based on the GCN and LSTM. EffHandEgoNet-Transformer [50], HTT [4], and Trear-depth [51] are based on CNN-Transformer architecture, in which EffHandEgoNet-Transformer [50] and HTT [4] use backbone based on CNN architecture for feature extraction.

In addition, several representative methods in the more challenging H2O dataset are also selected for comparison, including H+O [3], H2O [2], SlowFast [25], C2D [52], I3D [24], HTT [4], PoseConv3D [53]. These methods cover the current mainstream action recognition frameworks and can better reflect the overall state of the art. Tables 2 and 3 summarize the comparison results of action recognition accuracy on the FPHA dataset and H2O dataset, respectively. As can be seen from the tables, the methods proposed in this study achieve recognition performance better than existing methods on both datasets, which verifies its powerful ability to recognize the action of object interaction in complex scenes.

Table 2. Comparison of action recognition accuracy on FPHA set.

Method	Modality	Acc.
Joule-color-all [43]	RGB + Depth + Skeleton	78.78
Two stream-color [26]	RGB	61.56
Two stream-flow [26]	RGB	69.91
Two stream-all [26]	RGB	75.30
FPHA [1]	Pose	78.73
H+O [3]	RGB	82.43
HAN-2S [44]	Skeleton	89.04
Collaborative [5]	RGB	85.22
ResGCNeXt [45]	Skeleton	89.04
SBI-DHGR [46]	Skeleton	92.48
Li et al. [49]	RGB + Depth + Skeleton	91.95
HTT [4]	RGB	94.09
EffHandEgoNet-Transformer [50]	RGB	94.43
Trear-depth [51]	Depth	92.17
K. Prasse et al. [47]	Skeleton	92.52
SymNet-v2 [54]	RGB	82.96
GCN-BL [48]	Skeleton	80.52
Ours	RGB	**94.82**

Table 3. Comparison of action recognition accuracy on H2O set.

Method	Modality	Acc.
H2O w/ST-GCN [30]	Skeleton	73.86
H2O w/TA-GCN [2]	RGB + Depth	79.25
PoseConv3D [53]	Skeleton	83.47
H+O [3]	RGB	68.88
SlowFast [25]	RGB	77.69
C2D [52]	RGB	70.66
I3D [24]	RGB	75.21
HTT [4]	RGB	86.36
Ours	RGB	**87.92**

4.3.3. Ablation Studies

In this subsection, we will first perform an ablation study on the FPHA and H2O datasets to verify the effect of image features on action recognition accuracy. Table 4 reports the results of decimation experiments on these two datasets. Subsequently, the impact of different loss weights for parameters α_1, α_2, α_3 is explored on the FPHA dataset, as shown in Table 5.

Table 4. Ablation study of hand estimator in HPEM module with respect to Spatially Detailed Feature Extraction Part.

		w/o $f_{det}(t)$	w/$f_{det}(t)$
FPHA	Acc.	94.43	**94.82**
H20	Acc.	87.04	**87.92**
	hand.Left	35.26	**31.16**
	hand.Right	38.87	**35.21**

Table 5. A study on parameters α_3 on the FPHA dataset.

α_3	Acc.
0.5	92.47
0.75	93.88
1.0	**94.82**
1.25	94.09

We investigated the effect of the Spatially Detailed Feature Extraction Part on the accuracy of action recognition. Specifically, if only the global context features $f_{glo}(t)$ obtained from the transformer Encoder Block processing are used for action recognition, the model achieves 94.43% and 87.04% accuracy on the FPHA and H2O datasets, respectively, and the accuracy of pose estimation for both hands on the H2O dataset is 35.26 mm and 38.87 mm, respectively. These results show that modelling based on global contextual features $f_{glo}(t)$ is able to achieve relatively excellent action recognition performance, but there is still room for further optimization. To further enhance the model's ability to capture detailed spatial information, we added the Spatially Detailed Feature Extraction Part in the pose estimator for enhancing the perception of the local structure and spatial distribution of the hand. The Spatially Detailed Features $f_{det}(t)$ are concatenated and fused with the global contextual features $f_{glo}(t)$ extracted by the transformer Encoder Block through a dual residual network structure, resulting in a more expressive joint feature representation. The experimental results show that the inclusion of Spatially Detailed Features improves the model's action recognition accuracy by 0.39% and 0.88% on the two datasets and the left- and right-hand pose estimation accuracy by 4.1 mm and 3.66 mm on the H2O dataset, both of which achieve the best performance. This enhancement amply demonstrates that

our proposed design, i.e., Spatially Detailed Feature Extraction Part incorporating a dual residual network structure, is significantly effective in improving model performance. In addition, when the Spatially Detailed Feature Extraction Part is introduced, our frame parameter size is 113.8 MB, the inference time is 0.178 s, and we are able to achieve a frame rate of 5.65 FPS. This shows that our model performs better in balancing computational complexity as well as recognition accuracy.

We investigated the effect of different loss weights for the parameter α_3 on the recognition performance. The results in Table 5 show that our model reaches the highest accuracy when α_3 is set to 1.0, indicating that the current setting is the optimal configuration.

5. Conclusions

In this study, we propose a unified framework for recognizing hand actions and hand poses from first-person RGB videos. The framework effectively predicts the hand key point information and object category information for a single frame, and, at the same time, it is able to output the hand action categories in consecutive video frames. Specifically, HPEM estimates the hand key points in each image frame to obtain hand spatial information. Each video frame is first subjected to global feature extraction by FlexiViT and then fed to multiple morphing encoders to capture the relative positions of elements in the time-series data. Our hand pose estimator employs a dual residual structure to introduce a Spatially Detailed Feature Extraction Part, which uses MLP to predict hand key point coordinates and object categories. The hand pose results and object categories' information—along with the corresponding frame-level features—are then concatenated to fully exploit spatial information from each RGB frame. The Action Recognition Module (ARM) aggregates feature across consecutive frames to reconstruct the hand's motion trajectory, thereby capturing temporal dynamics over time. Finally, an action classifier determines the action class based on the fused spatiotemporal representations derived from the entire video sequence. We assess our method using two first-person hand action datasets, with experimental results demonstrating its strong performance on both the FPHA and H2O datasets. We focus on using the transformer architecture to address the problem of a unified framework for pose estimation and action recognition, including the introduction of Flexivit into our approach during the initial feature extraction phase. However, when it comes to multi-task joint learning as well as transformer architectures, the computational complexity of the model will be higher and face more computational pressure, whereas the sparse attention mechanism aims to significantly reduce the computational complexity and storage requirements by selectively focusing on key parts of the sequence to improve the efficiency and performance of the model when dealing with long sequential data, which we leave as future work.

Author Contributions: Conceptualization, J.Y., X.W. and Q.G.; methodology, J.Y., J.L. and H.P.; software, J.Y., Y.C. and X.W.; validation, H.P., Q.G., Y.C. and X.W.; formal analysis, J.Y., X.W. and Q.G.; investigation, J.Y., X.W. and Q.G.; resources, J.Y., X.W., Y.C. and Q.G.; data curation, J.Y., Q.G. and J.L.; writing—original draft preparation, J.Y.; writing—review and editing, Y.C., Q.G., H.P., J.L. and X.W.; visualization, J.Y.; supervision, Q.G. and X.W.; project administration, J.Y., X.W. and Q.G.; funding acquisition, Q.G. and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Natural Science Foundation of China (No. 62072362, 12101479), Natural Science Basis Research Plan in Shaanxi Province of China (Nos. 2021JQ-660 and 2024JC-YBMS-531), Shaanxi Provincial Innovation Capacity Support Programme Project (No. 2024ZC-KJXX-034), and Xi'an Major Scientific and Technological Achievements Transformation Industrialization Project (No. 23CGZHCYH0008).

Data Availability Statement: Here, we provide links to the two datasets used in our experiments. FPHA: https://guiggh.github.io/publications/first-person-hands/ (accessed on 11 September 2024); H2O: https://taeinkwon.com/projects/h2o/ (accessed on 4 July 2024).

Acknowledgments: We thank the editor and the anonymous reviewers for their valuable suggestions to improve the quality of this study. The feedback not only helped us to identify the deficiencies in the original manuscript but also provided clear directions on how to further improve and refine the study. We would like to express our sincerest gratitude to all those who have given us support, encouragement and provided valuable resources.

Conflicts of Interest: The authors declare that they have no known conflicting financial interests or personal relationships that could have appeared to influence the work reported in this study.

References

1. Garcia-Hernando, G.; Yuan, S.; Baek, S.; Kim, T.-K. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 409–419.
2. 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, Montreal, BC, Canada, 11–17 October 2021; pp. 10138–10148.
3. Tekin, B.; Bogo, F.; Pollefeys, M. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4511–4520.
4. 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, Vancouver, Canada, 18–22 June 2023; pp. 21243–21253.
5. Yang, S.; Liu, J.; Lu, S.; Er, M.H.; Kot, A.C. Collaborative learning of gesture recognition and 3D hand pose estimation with multi-order feature analysis. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part III 16, 2020. Springer: Berlin/Heidelberg, Germany, 2020; pp. 769–786.
6. Fan, Z.; Liu, J.; Wang, Y. Adaptive computationally efficient network for monocular 3d hand pose estimation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part IV 16, 2020. Springer: Berlin/Heidelberg, Germany, 2020; pp. 127–144.
7. Iqbal, U.; Molchanov, P.; Gall, T.B.J.; Kautz, J. Hand pose estimation via latent 2.5 d heatmap regression. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 18–22 June 2018; pp. 118–134.
8. Kim, D.U.; Kim, K.I.; Baek, S. End-to-end detection and pose estimation of two interacting hands. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11189–11198.
9. Lin, K.; Wang, L.; Liu, Z. Mesh graphormer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 12939–12948.
10. Meng, H.; Jin, S.; Liu, W.; Qian, C.; Lin, M.; Ouyang, W.; Luo, P. 3d interacting hand pose estimation by hand de-occlusion and removal. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2020; pp. 380–397.
11. Moon, G.; Yu, S.-I.; Wen, H.; Shiratori, T.; Lee, K.M. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part XX 16. 2020. Springer: Berlin/Heidelberg, Germany, 2020; pp. 548–564.
12. Mueller, F.; Bernard, F.; Sotnychenko, O.; Mehta, D.; Sridhar, S.; Casas, D.; Theobalt, C. Ganerated hands for real-time 3d hand tracking from monocular rgb. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 49–59.
13. Pan, H.; Cai, Y.; Yang, J.; Niu, S.; Gao, Q.; Wang, X. HandFI: Multilevel Interacting Hand Reconstruction Based on Multilevel Feature Fusion in RGB Images. *Sensors* **2024**, *25*, 88. [CrossRef] [PubMed]
14. Spurr, A.; Iqbal, U.; Molchanov, P.; Hilliges, O.; Kautz, J. Weakly supervised 3d hand pose estimation via biomechanical constraints. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 211–228.
15. Zimmermann, C.; Brox, T. Learning to estimate 3d hand pose from single rgb images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4903–4911.

16. Cai, Y.; Ge, L.; Liu, J.; Cai, J.; Cham, T.-J.; Yuan, J.; Thalmann, N.M. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2272–2281.
17. Chen, L.; Lin, S.-Y.; Xie, Y.; Lin, Y.-Y.; Xie, X. Temporal-aware self-supervised learning for 3d hand pose and mesh estimation in videos. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 1050–1059.
18. Wang, J.; Mueller, F.; Bernard, F.; Sorli, S.; Sotnychenko, O.; Qian, N.; Otaduy, M.A.; Casas, D.; Theobalt, C. Rgb2hands: Real-time tracking of 3d hand interactions from monocular rgb video. *ACM Trans. Graph. (ToG)* **2020**, *39*, 1–16. [CrossRef]
19. Cosma, A.; Radoi, E. GaitFormer: Learning Gait Representations with Noisy Multi-Task Learning. *arXiv* **2023**, arXiv:2310.19418.
20. Hu, L.; Gao, L.; Liu, Z.; Feng, W. Continuous sign language recognition with correlation network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 2529–2539.
21. Kim, J.-H.; Kim, N.; Won, C.S. Multi Modal Facial Expression Recognition with Transformer-Based Fusion Networks and Dynamic Sampling. *arXiv* **2023**, arXiv:2303.08419.
22. Xia, Z.; Peng, W.; Khor, H.-Q.; Feng, X.; Zhao, G. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 8590–8605. [CrossRef]
23. Zhu, X.; Huang, P.-Y.; Liang, J.; De Melo, C.M.; Hauptmann, A.G. Stmt: A spatial-temporal mesh transformer for mocap-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 1526–1536.
24. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
25. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6202–6211.
26. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
27. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal lstm with trust gates for 3d human action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2020; pp. 816–833.
28. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 12026–12035.
29. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 568–576.
30. 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, New Orleans, LA, USA, 2–7 February 2018.
31. 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, Santiago, Chile, 13–16 December 2015; pp. 4489–4497.
32. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
33. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
34. Lin, K.; Wang, L.; Liu, Z. End-to-end human pose and mesh reconstruction with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1954–1963.
35. Hampali, S.; Sarkar, S.D.; Rad, M.; Lepetit, V. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11090–11100.
36. Beyer, L.; Izmailov, P.; Kolesnikov, A.; Caron, M.; Kornblith, S.; Zhai, X.; Minderer, M.; Tschannen, M.; Alabdulmohsin, I.; Pavetic, F. Flexivit: One model for all patch sizes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 14496–14506.
37. Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; Liu, J. Human action recognition from various data modalities: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 3200–3225. [CrossRef]
38. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
39. Lee, J.; Toutanova, K. Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
40. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

41. Hasson, Y.; Tekin, B.; Bogo, F.; Laptev, I.; Pollefeys, M.; Schmid, C. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 571–580.
42. Aboukhadra, A.T.; Malik, J.; Elhayek, A.; Robertini, N.; Stricker, D. Thor-net: End-to-end graformer-based realistic two hands and object reconstruction with self-supervision. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 1001–1010.
43. Hu, J.-F.; Zheng, W.-S.; Lai, J.; Zhang, J. Jointly learning heterogeneous features for RGB-D activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 5344–5352.
44. Liu, J.; Wang, Y.; Xiang, S.; Pan, C. HAN: An efficient hierarchical self-attention network for skeleton-based gesture recognition. *Pattern Recognit.* **2025**, *162*, 111343. [CrossRef]
45. Peng, S.-H.; Tsai, P.-H. An efficient graph convolution network for skeleton-based dynamic hand gesture recognition. *IEEE Trans. Cogn. Dev. Syst.* **2023**, *15*, 2179–2189. [CrossRef]
46. Narayan, S.; Mazumdar, A.P.; Vipparthi, S.K. SBI-DHGR: Skeleton-based intelligent dynamic hand gestures recognition. *Expert Syst. Appl.* **2023**, *232*, 120735. [CrossRef]
47. Prasse, K.; Jung, S.; Zhou, Y.; Keuper, M. Local spherical harmonics improve skeleton-based hand action recognition. In Proceedings of the DAGM German Conference on Pattern Recognition, Heidelberg, Germany, 19–22 September 2023; Springer: Berlin/Heidelberg, Germany, 2020; pp. 67–82.
48. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 10–17 October 2021; pp. 13359–13368.
49. Li, R.; Wang, H. Graph convolutional networks and LSTM for first-person multimodal hand action recognition. *Mach. Vis. Appl.* **2022**, *33*, 84. [CrossRef]
50. Mucha, W.; Kampel, M. In my perspective, in my hands: Accurate egocentric 2d hand pose and action recognition. In Proceedings of the 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), Istanbul, Turkey, 27–31 May 2024; pp. 1–9.
51. Li, X.; Hou, Y.; Wang, P.; Gao, Z.; Xu, M.; Li, W. Trear: Transformer-based rgb-d egocentric action recognition. *IEEE Trans. Cogn. Dev. Syst.* **2021**, *14*, 246–252. [CrossRef]
52. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
53. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 2969–2978.
54. Wang, R.; Wu, X.-J.; Kittler, J. SymNet: A simple symmetric positive definite manifold deep learning method for image set classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 2208–2222. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.