Gated Temporal Shifts with Depth-Efficient Channel Attention for Real-Time Hand-Gesture Interaction

Salah-Eddine Laidoudi, Madjid Maidi, Samir Otmane

Gated Temporal Shifts with Depth-Efficient Channel Attention for Real-Time Hand-Gesture Interaction

Salah-eddine Laidoudi
ESME Research Lab / IBISC,
Université Paris-Saclay, Univ Evry
Paris, France
salah-eddine1.laidoudi@esme.fr

Madjid Maidi
LIASD, Université Paris 8
Saint-Denis, France
madjid.maidi@univ-paris8.fr

Samir Otmane
IBISC, Université Paris-Saclay,
Univ Evry
Évry, France
samir.otmane@univ-evry.fr

Abstract

We introduce a compact video-classification pipeline for real-time dynamic hand-gesture recognition in mixed-reality (MR) settings. The network marries a MobileNetV3 backbone with two purpose-built temporal components: (1) a Gated Discriminative Temporal Shift Module (G-DiTSM) that inserts first-order motion differences and learns channel-wise gates to fuse them adaptively, and (2) a lightweight Depth-Efficient Channel Attention (DepthECA) block that recalibrates spatial features on the fly. Operating on eight sparsely sampled frames per clip (Temporal Segment Network paradigm), the resulting model contains 2.65 M parameters and requires only 0.084 GFLOPs per inference. Evaluated on the RGB-only 20BN Jester benchmark (148k clips spanning 27 gesture classes) recorded from front-facing viewpoints. The system reaches 95.34% Top-1 and 99.80% Top-5 accuracy, surpassing recent 3D CNNs and transformer baselines while being an order of magnitude lighter. Ablations confirm that DepthECA and G-DiTSM provide complementary gains (+18.78% and +0.93% Top-1, respectively, over the MobileNetV3 baseline). Because all components are plug-and-play and introduce minimal overhead, the architecture is well suited to the tight latency and power budgets of standalone MR headsets, paving the way for natural grab, rotate, and command interactions using only on-board RGB cameras

CCS Concepts

• **Computing methodologies → Activity recognition and understanding**.

Keywords

Gesture Recognition, Mixed Reality, Lightweight Networks, Temporal Shift, Channel Attention

ACM Reference Format:
Salah-eddine Laidoudi, Madjid Maidi, and Samir Otmane. 2025. Gated Temporal Shifts with Depth-Efficient Channel Attention for Real-Time Hand-Gesture Interaction. In *31st ACM Symposium on Virtual Reality Software and Technology (VRST '25), November 12–14, 2025, Montreal, QC, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3756884.3765982

VRST '25, Montreal, QC, Canada

ACM ISBN 979-8-4007-2118-2/2025/11
https://doi.org/10.1145/3756884.3765982

1 Introduction

Mixed Reality (MR) seeks to blur the boundary between the physical and virtual worlds by embedding interactive 3D content seamlessly into our everyday environment. At the heart of any immersive MR system lies the ability to interact with virtual objects in a natural, intuitive manner. Traditional 2D interfaces fall short in spatial contexts, whereas 3D interaction—encompassing gestures, spatial navigation, and depth awareness—enables users to manipulate holograms and digital assets as if they were tangible. Among the many paradigms for 3D interaction, dynamic hand gesture recognition stands out for its direct alignment with innate human communication patterns and its minimal hardware requirements: modern MR headsets already provide the necessary RGB camera streams, making gesture-driven control both scalable and user-friendly.

We introduce an RGB-only gesture-recognition pipeline that delivers state-of-the-art accuracy at mobile-class cost. The design starts from a MobileNetV3 backbone and adds two novel modules:

First, we propose a *Gated Discriminative Temporal Shift Module* (G-DiTSM) that extends the parameter-free Temporal Shift operator with learnable channel gates and a depth-wise 3D aggregation kernel, thereby injecting first-order motion information without costly 3D convolutions (Sec. 4.4). Second, we introduce a *Depth-Efficient Channel Attention* (DepthECA) block that combines dual global descriptors with a cross-stage reinforcement gate to recalibrate feature channels on the fly; the block adds only ∼0.002 M parameters yet markedly boosts accuracy (Sec. 4.2).

Use case & scope. Our RGB-only recognizer complements, rather than replaces, vendor hand tracking. On HMDs with native tracking (e.g., Quest 3), it serves as a *customizable dynamic-gesture command layer* that can be trained for task-specific gestures without OS integration. It also targets devices that lack native tracking (phones, tablets, camera-only smart glasses) and provides a purely-RGB fallback when IR/depth sensing degrades (e.g., bright sunlight, specular surfaces, partial occlusions).

To train and evaluate our system, we leverage the 20BN-Jester dataset, a comprehensive benchmark of over 148 000 video clips spanning 27 distinct hand gestures *captured by front-facing cameras (user-facing)* rather than HMD-peripheral viewpoints; we therefore treat Jester as a realistic, non-egocentric proxy for MR command gestures and discuss this domain gap in our limitations. Jester's focus on concise, well-annotated upper-body and hand motions makes it an ideal testbed for MR-centric gesture recognition, where swift and reliable motion understanding is critical. The resulting MobileNetV3 + TSM + DepthECA model achieves a harmonious balance of accuracy, robustness to complex backgrounds, and real-time inference speed—paving the way for natural 3D interactions

such as grabbing, rotating, scaling, and commanding virtual objects through simple hand gestures on resource-constrained MR headsets.

2 Dataset

2.1 20BN-Jester Dataset

The development of effective real-time mixed reality (MR) applications for dynamic hand gesture recognition depends heavily on the availability of suitable large-scale datasets. However, many existing gesture recognition datasets rely on depth sensors, infrared imaging, or complex multi-modal sensor configurations, which impose hardware requirements not always compatible with lightweight MR headsets. The **20BN-Jester dataset** [14] addresses these limitations by offering a purely RGB-based gesture corpus captured by front-facing cameras (user-facing). While widely used in MR research, Jester is not recorded from HMD-peripheral viewpoints.

The Jester dataset encompasses **27 distinct gesture classes**, covering a broad range of both dynamic (motion-based) and static (pose-based) hand gestures. Each video clip captures a single gesture instance with durations varying from 2 to 6 seconds. Recordings are *front-facing* (a fixed camera facing the user), not peripheral HMD views. We therefore treat Jester as a realistic proxy for MR command gestures while noting the viewpoint domain gap. The dataset is accompanied by high-quality annotations and predefined training, validation, and test splits, simplifying the evaluation and reproducibility of gesture recognition experiments.

Dataset Statistics and Composition. In total, the 20BN-Jester dataset comprises **148,092** video clips, with **118,562** for training, **14,787** for validation, and **14,743** for testing. The class distribution is shown in Fig. 1, and sample frames are shown in Fig. 2. Each clip represents a single labeled gesture. The average clip duration is approximately **3.7 seconds**, captured at a standard frame rate of **30 frames per second (FPS)**, resulting in roughly 112 frames per clip. For storage and computational efficiency, videos have been resized to a compact spatial resolution of **112×112 pixels**. Importantly, the dataset employs a subject-independent split protocol to ensure that training and evaluation sets do not contain overlapping participants, thereby preserving strict generalization criteria.

Comparative Context. To situate the characteristics of 20BN-Jester within the broader landscape of gesture recognition benchmarks, Table 1 summarizes key properties of Jester alongside several widely-used RGB and RGB-D gesture datasets. While multi-modal datasets such as **EgoGesture** [34], **NVGesture** [16], and **ChaLearn IsoGD** [27] leverage depth or multimodal sensors to enhance spatial representation, Jester remains strictly RGB-based, eliminating dependency on additional hardware and better reflecting real-world deployment constraints for mobile MR systems. Compared to larger-scale video action datasets such as **Something-Something V2** [8], Jester remains highly focused on *fine-grained hand gestures*, making it particularly suitable for interactive MR interfaces where hand-based control dominates.

Evaluation Protocol. Performance evaluation on Jester follows standard classification metrics widely adopted in gesture recognition tasks. The **Top-1 accuracy** reflects the percentage of video clips for which the predicted label matches the ground truth. The **Top-5 accuracy** considers whether the true label appears among the top five model predictions, providing additional insight into model uncertainty. In addition to classification accuracy, practical MR deployment scenarios require reporting of **inference speed**, typically measured in frames-per-second (FPS), to ensure that real-time interaction requirements are met.

Practical Considerations for MR Applications. The 20BN-Jester dataset offers several properties that render it particularly well-suited for developing gesture recognition pipelines targeting mixed reality systems. Temporal subsampling—where a fixed number of frames are sampled uniformly across the video duration—reduces computational demands while preserving temporal dynamics. Combined with standard spatial augmentations such as random cropping, horizontal flipping, and color jittering, these augmentation strategies promote model robustness to variations in lighting, background, and camera viewpoint that commonly arise in unconstrained MR environments. Moreover, the short clip lengths and modest resolution facilitate rapid prototyping and enable lightweight deep models to meet the low-latency requirements necessary for immersive interaction. Finally, the dataset is publicly available for research use under the provider's license (registration required), which has encouraged reproducibility and broad adoption.

3 Related Work

Dynamic hand-gesture recognition has evolved through diverse paradigms—skeleton-based, RGB, depth, multi-modal fusion, and temporal modeling—each with distinct accuracy-latency-hardware trade-offs (Fig. 3). We position our **MobileNetV3-TSM-DepthECA** pipeline within this landscape, addressing limitations in prior approaches.

3.1 Skeleton-Based Methods

These approaches leverage estimated hand keypoints for trajectory analysis. Early geometric descriptors like Fisher vectors over joint positions [5] offered lighting robustness but required precise pose estimation. Recent graph-based methods model spatio-temporal relationships through GCNs: ST-GCN [32] treats joints as graph nodes, while adaptive variants (Two-Stream AGCN [20], Shift-GCN [4]) learn dynamic adjacency matrices. Though efficient due to low-dimensional inputs, these methods degrade with tracking noise and lose fine finger details when only joints are considered.

3.2 RGB-Based Methods

Vision-based approaches dominate with several architectural innovations:

- *Two-stream CNNs*: Fuse RGB appearance and optical flow motion streams [21] at computational cost
- *3D CNNs*: Learn spatio-temporal features directly (C3D [24], I3D [3]) but incur heavy parameterization
- *Efficient hybrids*: TSN samples sparse frames for consensus prediction [28], while TSM enables temporal modeling via feature shifting [12]. Decomposed 3D convolutions (P3D [19], R(2+1)D [26]) reduce complexity

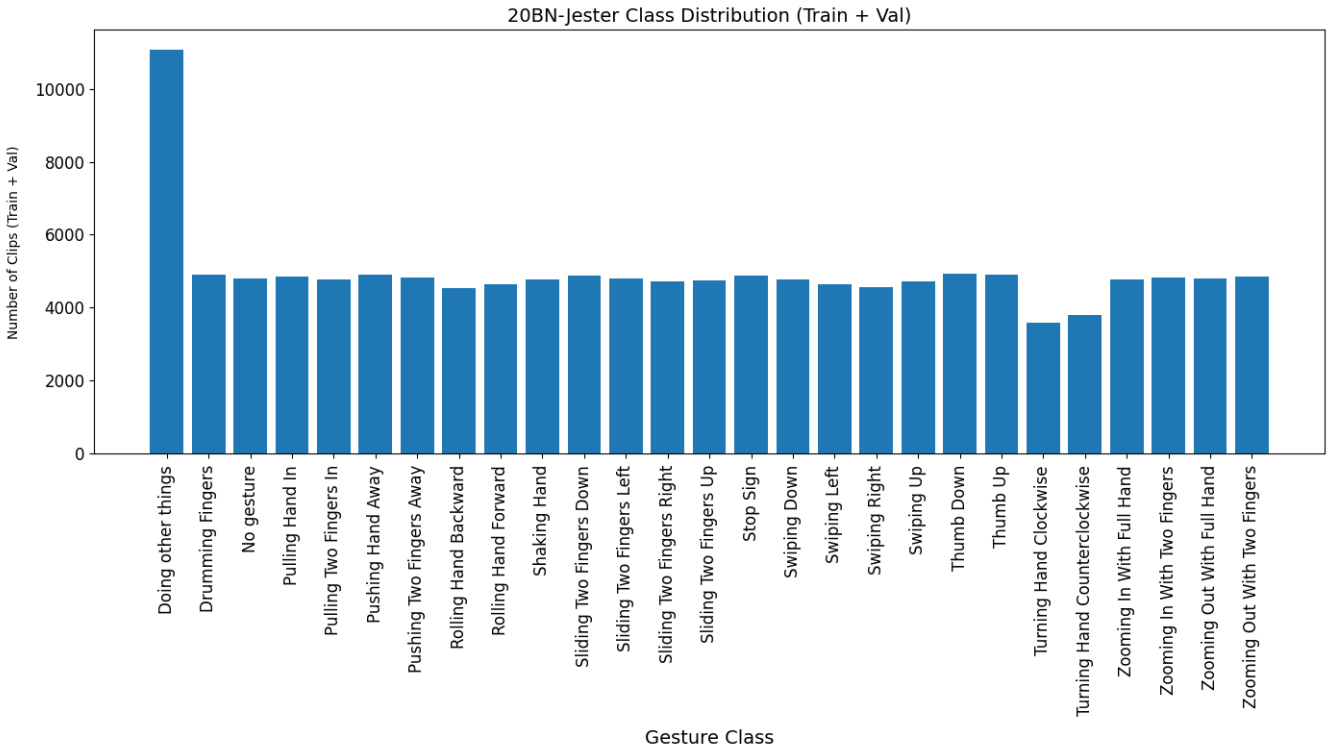

Figure 1: Class frequency distribution across the training and validation splits of the Jester dataset.

Table 1: Comparison between Jester and representative gesture recognition datasets.

Dataset	Modality	#Classes	#Clips	Avg. Length	Egocentric	Depth
20BN-Jester	RGB	27	148,092	3.7 s	No (front-facing)	No
EgoGesture	RGB-D	83	24,161	1.5 s	Yes	Yes
NVGesture	RGB-D	25	1,377	2.1 s	No	Yes
ChaLearn IsoGD	RGB-D	249	47,933	2.6 s	No	Yes
Something-Something V2	RGB	174	220,847	3.3 s	No	No

- *Transformers*: Capture long-range dependencies (TimeSformer [2], ViViT [1]) but require substantial resources

RGB methods capture rich appearance cues but remain sensitive to lighting/background changes.

3.3 Depth-Based Methods

Depth sensors provide lighting-invariant 3D structure. Beyond hand-crafted features (HON4D, super-normal vectors), deep approaches include 3D CNN-RNN hybrids [16] that enable online segmentation. Point-cloud methods like PointLSTM [15] process sampled 3D points via spatio-temporal networks. While effective in low-light conditions, they lack color/texture cues and depend on sensor quality.

3.4 Multi-Modal Fusion

Sensor fusion improves robustness against modality-specific failures. Neverova *et al.* [17] pioneered three-stream fusion (RGB, depth, skeleton) with ModDrop training. Molchanov *et al.* [16] combined color/depth/IR for automotive applications. Recent adaptive methods dynamically weight modalities but increase system complexity and require synchronized sensors.

3.5 Temporal Modeling

Explicit sequence handling is crucial for dynamic gestures:

- *RNNs/LSTMs*: Model variable-length sequences [6, 18] but suffer sequential processing latency
- *Temporal CNNs*: Use 1D convolutions for parallel computation [11], potentially enhanced with dilations
- *Attention mechanisms*: Capture long-range dependencies efficiently (Non-local Nets [30]) or via transformers [36]

Figure 2: Sample frames illustrating several gesture categories from 20BN-Jester (front-facing camera, not HMD-peripheral).

Hybrid approaches (e.g., CNN+LSTM) remain popular for joint spatial-temporal learning.

3.6 Lightweight Models

Deployable systems require efficiency optimizations:

- *Backbones*: MobileNetV3 [9] (depthwise separable convolutions), ShuffleNet [35] (channel shuffling), EfficientNet [22] (balanced scaling)
- *Temporal modules*: TSM [12] (zero-parameter shifting), X3D [7] (progressive 3D expansion)
- *Compression*: Pruning, quantization, and distillation for acceleration

These achieve real-time performance but risk accuracy loss on fine-grained gestures.

3.7 Our Approach

We bridge critical gaps in prior work by:

(1) Adopting MobileNetV3's efficient architecture while processing raw RGB frames (avoiding skeleton dependency);
(2) Enhancing TSM with **Gated-DiTSM** for content-adaptive temporal fusion;
(3) Introducing **Depth-Efficient Channel Attention (DepthECA)** to amplify discriminative *channel* features at minimal overhead.

Our design avoids heavy 3D convolutions/transformers while outperforming skeleton-based methods on fine finger motions.

4 Methodology

In this section, we present a detailed description of our dynamic hand gesture recognition pipeline that is illustrated in Figure 4, carefully explaining both the architectural design choices and their mathematical underpinnings. Our proposed system combines efficient spatial feature extraction with novel temporal modeling and adaptive attention mechanisms, striking a balance between accuracy and real-time performance suitable for resource-constrained mixed reality (MR) headsets.

We first describe the Temporal Segment Network (TSN) framework as our high-level temporal abstraction backbone (Section 4.1), then introduce the Discriminative Temporal Shift Module (DiTSM) to capture local frame differences (Section 4.3), followed by our novel Gated DiTSM (G-DiTSM), which adds content-adaptive temporal fusion (Section 4.4), and finally provide an integration and complexity analysis (Section 4.5).

4.1 Temporal Segment Network Backbone

The starting point of our pipeline is the well-established Temporal Segment Network (TSN) framework [28], which provides a simple yet effective way to model long-term temporal dependencies in video by sampling frames sparsely across the entire video duration. This allows us to reduce computational cost without sacrificing temporal coverage.

Given a video clip with L frames, we divide it into T equal temporal segments. From each segment, one frame is randomly sampled for training (or center-sampled for inference), yielding T frames per clip. For a batch of N clips, this produces:

$$\text{Sampled frames: } F_{n,t} \in \mathbb{R}^{3\times H_0\times W_0}, \quad n = 1\dots N,\ t = 1\dots T.$$

After standard preprocessing—spatial resizing, random cropping, normalization, and optional horizontal flipping—we stack all $N \times T$ frames into a single 4D input tensor:

$$X \in \mathbb{R}^{(N\cdot T)\times 3\times H\times W}.$$

Lightweight 2D Backbone for Per-Frame Feature Extraction. We adopt MobileNetV3 as our core 2D CNN backbone [9]. This choice is motivated by its excellent trade-off between accuracy and efficiency, achieved through depthwise separable convolutions, squeeze-and-excitation (SE) attention modules, and architecture search. The backbone processes the input tensor frame-wise, producing:

$$F = \mathcal{B}(X) \in \mathbb{R}^{(N\cdot T)\times C_{\text{feat}}}.$$

Here, C_{feat} is the feature dimension of the final embedding for each frame.

Classification Head and Temporal Consensus. After feature extraction, we apply dropout regularization and a fully connected layer for classification:

$$Z = \text{Dropout}(F; p)\cdot W_{\text{fc}} + b_{\text{fc}} \in \mathbb{R}^{(N\cdot T)\times K}.$$

The outputs are reshaped back into $\mathbb{R}^{N\times T\times K}$ for temporal aggregation. TSN employs simple mean-consensus across the T segments:

$$\hat{y}_{n,k} = \frac{1}{T}\sum_{t=1}^{T} Z_{n,t,k}, \quad \forall n, k.$$

Finally, softmax produces class probabilities. This design allows TSN to efficiently integrate long-range temporal information with minimal overhead, but it lacks fine-grained temporal modeling of short-term motion patterns between frames. We address this next.

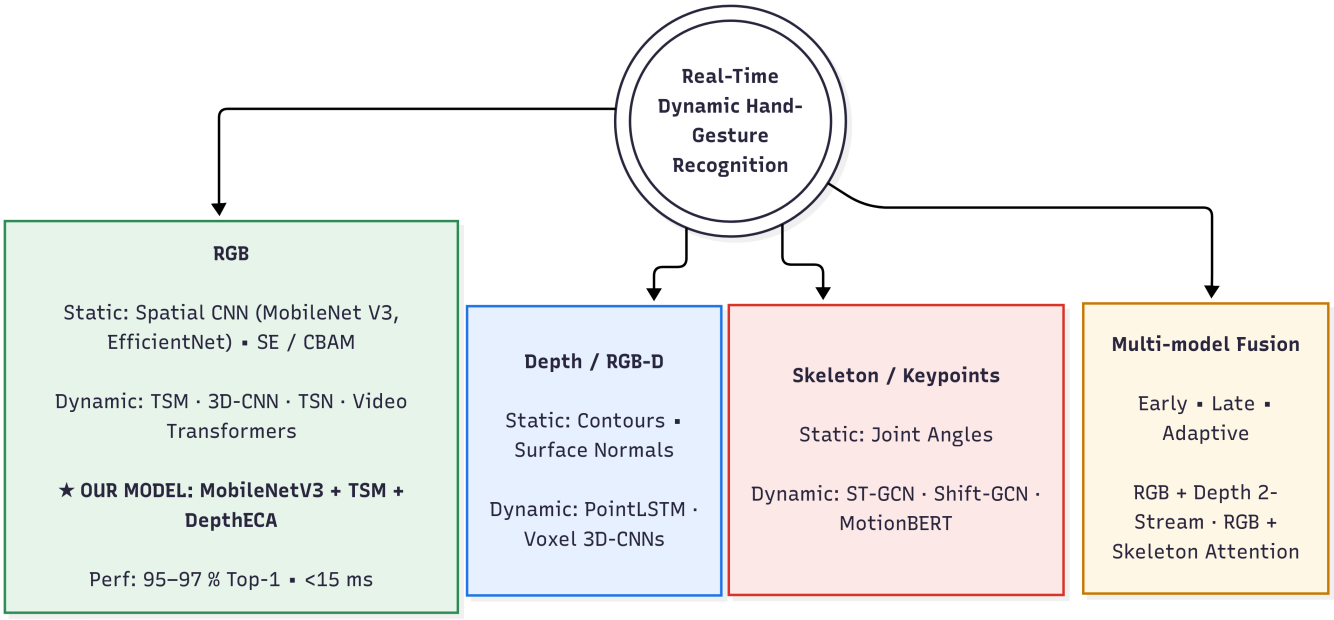

Figure 3: Comprehensive Topology of Real-Time Dynamic Hand-Gesture Recognition Techniques—Covering RGB, Depth, Keypoint, and Fusion Modalities with CNN, GCN, Transformer, and Attention Mechanisms

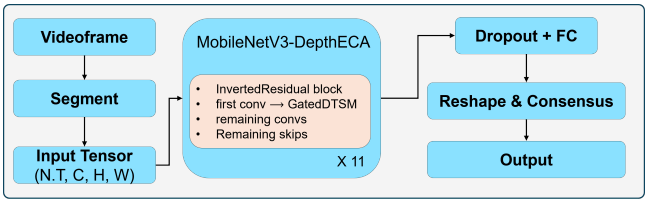

Figure 4: Overview of our TSN-based architecture enhanced with DiTSM and Gated-DiTSM modules, integrated into the MobileNetV3 backbone.

4.2 Depth-Efficient Channel Attention (DepthECA)

DepthECA is a lightweight channel–recalibration block that extends the squeeze-and-excite family. Given a spatial feature tensor $X_{\text{spatial}} \in \mathbb{R}^{B\times C\times H\times W}$, we derive two complementary global descriptors via adaptive average and max pooling,

$$X^c_{\text{avg}} = \text{AdaptiveAvgPool}(X_{\text{spatial}}), \tag{1}$$

$$X^c_{\text{max}} = \text{AdaptiveMaxPool}(X_{\text{spatial}}). \tag{2}$$

The two C-dimensional vectors are concatenated and processed by a SiLU-activated bottleneck of two 1×1 convolutions. In parallel, a *cross-stage reinforcement* branch applies an additional 1×1 convolution to the current feature map; its output is scaled by a learnable scalar γ (initialised to 0.1). The resulting channel-attention weights are

$$w = \sigma\Big(\underbrace{\text{Conv}_{1\times1}(\text{SiLU}(\text{Conv}_{1\times1}[X^c_{\text{avg}} \| X^c_{\text{max}}]))}_{\text{dual-pool bottleneck}} + \gamma\ \text{Conv}^{\text{boost}}_{1\times1}(X_{\text{spatial}})\Big), \tag{3}$$

where σ denotes the sigmoid function. The attention vector $w \in \mathbb{R}^{B\times C\times1\times1}$ rescales the input channels,

$$X_{\text{channel}} = X_{\text{spatial}} \odot w. \tag{4}$$

DepthECA introduces only ~ 0.002 M additional parameters yet improves Top-1 accuracy on 20BN-Jester by +4.6% over the MobileNetV3 baseline (see Sec. 5.2). The overall dataflow is illustrated in Figure 5.

4.3 Discriminative Temporal Shift Module (DiTSM)

While TSN models global temporal structure, it does not capture local frame-to-frame motion explicitly. Motivated by this, we insert the Discriminative Temporal Shift Module (DiTSM) to introduce first-order temporal differences directly into the feature stream.

Tensor Reshaping for Segment-Level Processing. The input tensor $X \in \mathbb{R}^{(NT)\times C\times H\times W}$ is first reshaped into:

$$\widetilde{X} \in \mathbb{R}^{N\times T\times C\times H\times W}.$$

We then partition the feature channels into three groups based on the shift division factor $n_{\div}$:

$$\text{fold size} = \left\lfloor \frac{C}{n_{\div}} \right\rfloor.$$

The channels are allocated as:

- Forward shift: first fold channels capture forward differences.

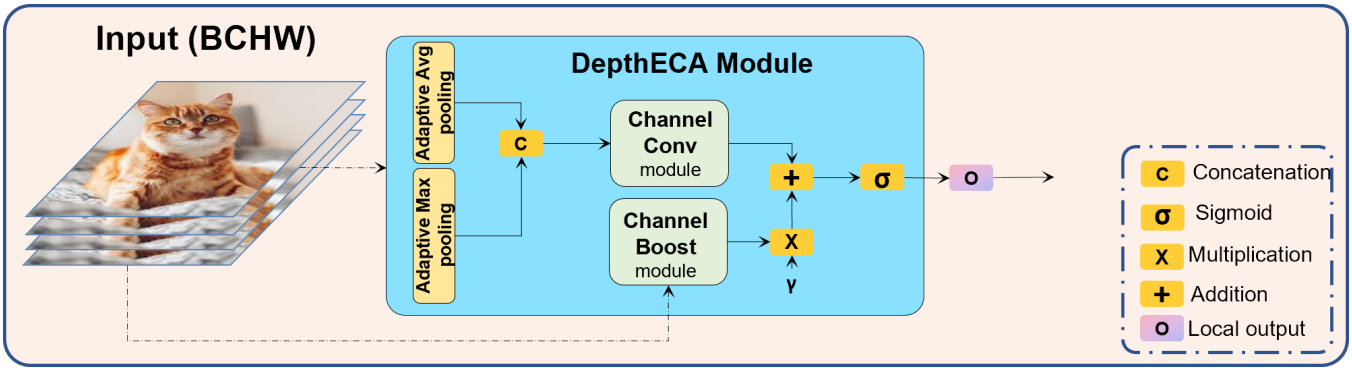

Figure 5: DepthECA pipeline. Dual global pooling produces complementary descriptors that flow through a 1×1 bottleneck (Channel Conv). A cross-stage residual (+) is modulated by a learnable scalar γ (Channel Boost). The sum is passed through sigmoid and applied to the input tensor to obtain the refined output O.

- Backward shift: next fold channels capture backward differences.
- Static: remaining channels remain unchanged.

Temporal Differences. We compute:

$$Y_{n,t,0:\text{fold},h,w} = \widetilde{X}_{n,t+1,0:\text{fold},h,w} - \widetilde{X}_{n,t,0:\text{fold},h,w}, \tag{5}$$

$$Y_{n,t,\text{fold:2fold},h,w} = \widetilde{X}_{n,t-1,\text{fold:2fold},h,w} - \widetilde{X}_{n,t,\text{fold:2fold},h,w}, \tag{6}$$

$$Y_{n,t,2\text{fold}:C,h,w} = \widetilde{X}_{n,t,2\text{fold}:C,h,w}. \tag{7}$$

This design allows DiTSM to inject explicit frame-to-frame motion differences while preserving much of the static appearance content, which is crucial for recognizing gestures that combine both pose and movement patterns.

Limitation of DiTSM.. While DiTSM introduces motion cues efficiently, its channel splitting is fixed, and it applies uniform differencing regardless of the input content. To improve upon this, we propose a content-adaptive gated extension.

4.4 Gated Discriminative Temporal Shift Module (G-DiTSM)

We introduce the Gated DiTSM (G-DiTSM) module, which extends DiTSM by integrating two learnable enhancements: temporal aggregation via depth-wise 3D convolution, and content-adaptive gating via SE-style attention (see Fig. 6).

Step 1 — Depthwise Temporal Aggregation. After DiTSM, we obtain tensor $Y \in \mathbb{R}^{N\times T\times C\times H\times W}$, which we permute into channel-first order for temporal convolution:

$$Y' \in \mathbb{R}^{N\times C\times T\times H\times W}.$$

We apply a depthwise 3D convolution:

$$\widetilde{Y}' = \text{BN}_3\left(\text{Conv3D}_{\text{dw}}(Y')\right),$$

with kernel size $(3, 1, 1)$, group size equal to the number of channels, and zero-initialized weights so the module starts with identity behavior.

This operation allows each channel to aggregate neighboring temporal differences, effectively modeling small temporal patterns directly.

Step 2 — SE-Style Channel Gating. Next, we adaptively weight feature channels using a squeeze-and-excitation mechanism:

$$s_n[c] = \frac{1}{THW}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W}\widetilde{Y}'_{n,c,t,h,w}.$$

A lightweight two-layer excitation network then computes:

$$u_n = \text{ReLU}(W_1 s_n), \quad z_n = W_2 u_n + b_2, \quad g_n = \sigma(z_n),$$

where $W_1 \in \mathbb{R}^{(C/r)\times C}$, $W_2 \in \mathbb{R}^{C\times(C/r)}$, and b_2 is initialized to -3 to bias gating towards initial suppression.

The gates g_n are broadcasted and reshaped to match the original tensor dimensions:

$$G \in \mathbb{R}^{N\times 1\times C\times 1\times 1}.$$

Step 3 — Gated Residual Fusion. Finally, we fuse the gated temporal conv output with the original shifted tensor via residual connection:

$$O = Y + G \odot \text{PermuteInv}(\widetilde{Y}').$$

The result is flattened back into $\mathbb{R}^{(NT)\times C\times H\times W}$, seamlessly feeding into the rest of the CNN backbone.

4.5 Integration Strategy and Computational Complexity

We inject G-DiTSM into the MobileNetV3 backbone by replacing the *first convolution layer* inside each inverted residual block. This ensures early temporal modeling while retaining efficient feature extraction in later stages.

Complexity Analysis. The computational overhead introduced by G-DiTSM is modest:

- **DiTSM shifts:** negligible (indexing only).
- **Depthwise Conv3D:** $3 \times H' \times W' \times C$ FLOPs per time step.
- **SE gating:** $2C^2/r$ multiplications (very small for typical reduction factor $r = 4$).

Overall, the added overhead is typically less than 5% of the total backbone computation—significantly smaller than 3D CNNs or full spatio-temporal transformers.

This design choice maintains the core philosophy of TSN: lightweight temporal modeling with minimal complexity, while our

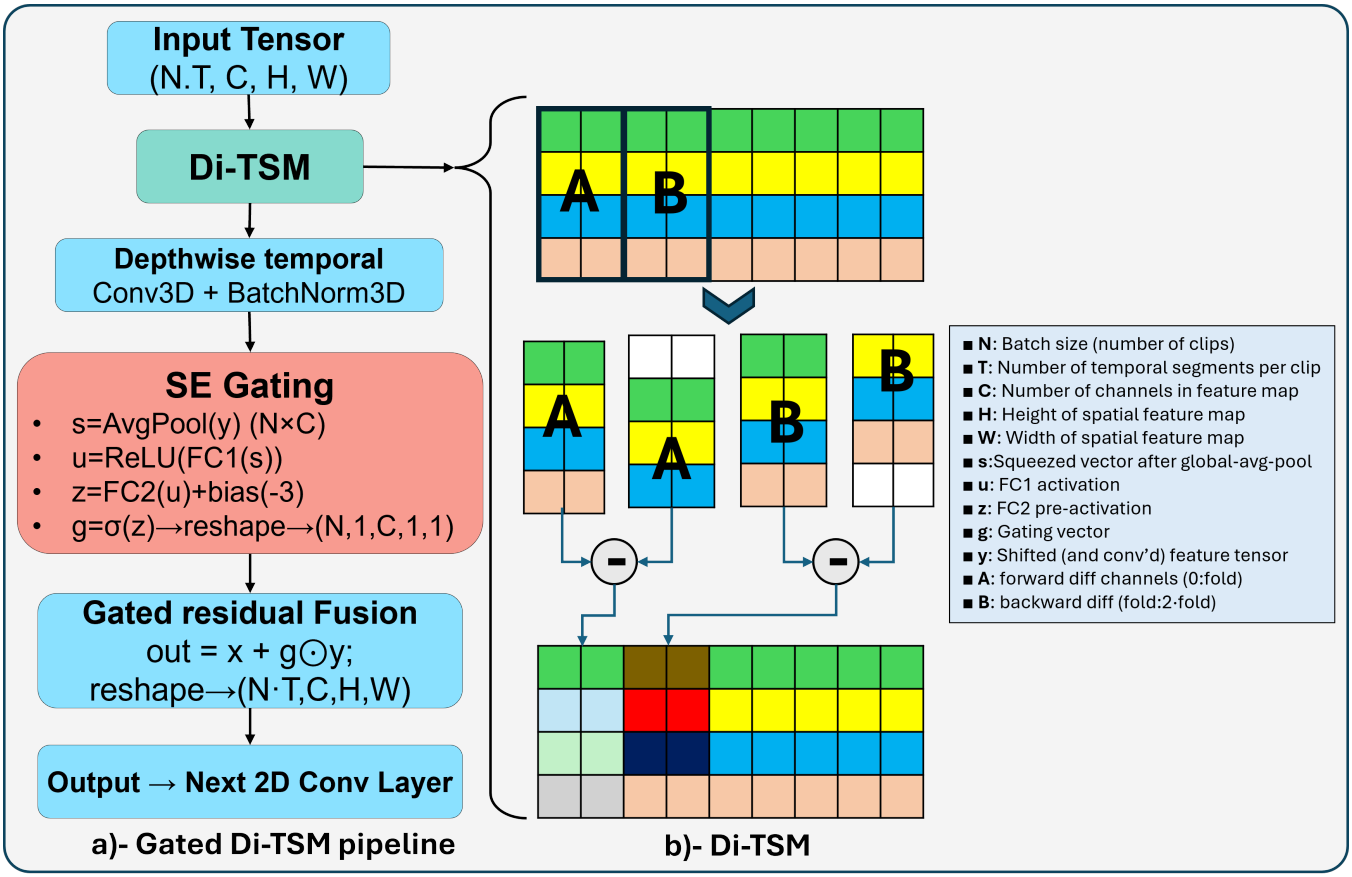

Figure 6: Internal structure of the proposed *Gated Discriminative TSM* (G-DiTSM). (a) End-to-end data-flow from the input tensor to the gated residual output. (b) Detail of the underlying DiTSM channel-shift operation that produces forward (A) and backward (B) difference channels. The legend (right) lists all tensor symbols used in Sec. 4.4.

gated temporal fusion captures richer motion cues than classical temporal pooling or pure shift-based modules.

Link to Experimental Evaluation. In Section 5, we empirically validate the contributions of each module through comprehensive ablation studies, showing the cumulative improvements of DepthECA, DiTSM, and G-DiTSM on top of the MobileNetV3 baseline.

5 Results and Discussion

We now present a comprehensive evaluation of our proposed method on the 20BN-Jester dataset, comparing it both to existing state-of-the-art gesture recognition models and through controlled ablation studies that isolate the contribution of each architectural component.

5.1 Comparison to State-of-the-Art

Table 2 reports performance of our model alongside several peer-reviewed gesture recognition methods evaluated on Jester. Our proposed model achieves a Top-1 accuracy of 95.34%, outperforming prior state-of-the-art approaches including ACTION-Net [31], ESTI [10], and STASTA [23], while maintaining a significantly lower computational footprint.

Compared to methods such as R3D, I3D, and ViViT-L (which employ full 3D convolutions or transformers), our model offers substantial efficiency gains with far fewer parameters, lower FLOPs, and smaller memory footprint — key factors for real-time mixed reality deployment. The combination of MobileNetV3, DiTSM, and Gated-DiTSM modules yields both superior accuracy and efficient inference.

5.2 Ablation Study

To rigorously analyze the individual contributions of each architectural component in our proposed model, we conduct an ablation study using the Jester dataset. The experiments follow a controlled incremental design, wherein modules are introduced progressively on top of a common baseline to isolate their respective effects on recognition performance, parameter count, computational cost, and memory usage. The results are summarized in Table 3.

We begin with the **MobileNetV3 baseline**, which provides a compact and efficient 2D convolutional backbone operating solely on per-frame information. As expected, this frame-wise model

Table 2: Performance comparison on 20BN-Jester benchmark (ranked by Top-1 accuracy).

Model	Top-1 Acc (%)	Top-5 Acc (%)	Params (M)	FLOPs (G)	Mem (MB)
Ours	**95.34**	**99.80**	2.65	0.084	7.9
ESTI [10]	94.47	99.63	27.11	38.00	101
ACTION-Net [31]	93.53	99.55	2.28	15.75	12
STASTA [23]	92.62	99.49	24.80	48.16	95
CvT-MTAM [13]	92.18	—	24.60	64.52	98
PAN [33]	91.51	—	28.40	30.62	112
R3D (3D ResNet) [25]	90.82	—	44.45	84.33	174
R(2+1)D-34 [26]	90.20	—	44.10	82.74	162
TSM [12]	89.80	96.42	23.56	39.66	91
ViViT-L (Transformer) [1]	81.70	93.80	31.80	144.00	120
TSN [29]	81.00	99.00	23.68	33.00	94

yields limited temporal modeling capacity, achieving a Top-1 accuracy of 75.54% and Top-5 accuracy of 97.80%.

Next, we introduce the **DepthECA** module, which recalibrates channel-wise feature responses through lightweight channel attention. This addition yields a notable improvement of nearly 5% absolute Top-1 accuracy, confirming the importance of dynamically weighting salient channels even within purely spatial representations.

The subsequent integration of the **Temporal Shift Module (TSM)** introduces an implicit temporal modeling capability by allowing information exchange across frames via channel shifts. This lightweight operation dramatically boosts performance, improving Top-1 accuracy to 93.75%, demonstrating the crucial role of temporal context for dynamic gesture understanding.

To further enhance temporal sensitivity, we replace the basic shift mechanism with our proposed **Gated Discriminative Temporal Shift Module (G-DiTSM)**. The incorporation of content-adaptive gating and learnable temporal aggregation delivers an additional accuracy gain to 94.41%, while incurring only marginal computational overhead. This result validates the effectiveness of gating mechanisms in refining shifted temporal features beyond simple channel reassignment.

Finally, combining **DepthECA** with **Gated-DiTSM** yields our full model configuration. The joint modeling of channel recalibration and gated temporal shifts achieves the best overall performance, attaining a Top-1 accuracy of 95.34% and Top-5 accuracy of 99.80%, while maintaining an extremely compact parameter count of 2.65M and computational cost of only 0.084 GFLOPs. These findings empirically confirm the complementary nature of the two modules, each addressing different aspects of spatio-temporal representation.

5.3 Discussion

The ablation results presented above provide several important insights into the behavior of our proposed architecture and the relative contribution of each component.

First, we observe that **temporal modeling is indispensable** for dynamic hand gesture recognition. The introduction of the Temporal Shift Module (TSM) yields the most significant performance gain relative to the baseline, confirming that per-frame spatial representations are insufficient to capture motion patterns inherent to dynamic gestures. The ability of TSM to inject temporal context without introducing additional parameters or convolutions highlights its exceptional efficiency for real-time video applications.

Second, while TSM provides a strong temporal backbone, its static channel shifting design can still benefit from further refinement. The proposed **Gated Discriminative Temporal Shift Module (G-DiTSM)** introduces content-adaptive gating, which allows the network to selectively modulate the contribution of temporal differences on a per-channel basis. This adaptive fusion mechanism leads to consistent improvements over the rigid shifts of TSM, demonstrating that not all motion features contribute equally to gesture discrimination, and that learning to weight them dynamically is beneficial.

Third, **channel-wise recalibration** via the DepthECA module plays a complementary role by enhancing feature selectivity at the spatial level. Even without temporal modeling, DepthECA alone improves baseline accuracy, suggesting that recalibrating spatial feature responses is valuable for capturing subtle hand shape variations across frames. When combined with temporal modules, DepthECA consistently contributes additional accuracy gains, confirming its utility in balancing spatial and temporal feature emphasis.

Moreover, the complete model preserves a highly favorable **efficiency-accuracy trade-off**. Despite achieving Top-1 accuracy exceeding 95%, the model maintains a parameter count of only 2.65M and computational cost below 0.1 GFLOPs, far surpassing several prior state-of-the-art methods in both recognition performance and computational economy (cf. Table 2). This lightweight profile enables real-time deployment on resource-constrained mixed reality hardware without sacrificing accuracy.

Finally, it is noteworthy that the modular nature of our design allows each component to be integrated independently or jointly depending on available computational budget or hardware constraints. The architecture generalizes naturally across a spectrum of deployment scenarios — from ultra-lightweight configurations (e.g., DepthECA + TSM for mobile inference) to higher-performance variants (e.g., DepthECA + G-DiTSM for high-accuracy systems).

Qualitative results (Fig. 7) show stable predictions across common command gestures.

In summary, the proposed integration of temporal shifts, gated attention, and channel recalibration offers a robust and flexible

Table 3: Ablation study results on the Jester dataset (8-frame input setting).

Variant	Top-1 (%)	Top-5 (%)	Params (M)	FLOPs (G)	Mem (MB)
Baseline (MobileNetV3)	75.54	97.80	2.50	0.06	9.8
+ DepthECA	80.12	98.62	2.55	0.07	7.9
+ TSM	93.75	99.28	2.50	0.06	9.8
+ Gated-DiTSM	94.41	99.67	2.60	0.08	7.8
+ DepthECA + TSM	94.78	99.72	2.60	0.08	7.9
+ DepthECA + Gated-DiTSM (Ours)	**95.34**	**99.80**	2.65	0.084	7.9

Figure 7: Qualitative hand-gesture inference examples in our RGB-only MR prototype. Predicted class and per-frame confidence (top-1 softmax) are overlaid in cyan. Gestures: (a) Drumming Fingers, (b) Sliding Two Fingers Down, (c) Swiping Left, (d) Swiping Right, (e) Zooming In With Two Fingers, (f) Thumb Up.Yellow text shows the end-to-end camera loop rate on our prototype laptop; this is not XR-SoC throughput.

solution for real-time dynamic hand gesture recognition, capable of adapting to the computational requirements and accuracy demands of next-generation mixed reality interaction systems.

6 Conclusion

This paper presented a complete, real-time pipeline for RGB-only hand-gesture recognition on untethered MR headsets. Building on a MobileNetV3 backbone, we contributed two lightweight yet complementary modules: the *Gated Discriminative Temporal Shift*

Module (G-DiTSM), which injects content-aware first-order motion at the cost of a 2-D network, and the *Depth-Efficient Channel Attention* (DepthECA) block, which recalibrates feature channels with only ∼0.002 M additional parameters. Combined, these components enable the model to attain **95.3 % Top-1** and **99.8 % Top-5** accuracy on the 20BN-Jester benchmark while requiring just **2.65 M** parameters and **0.084 GFLOPs**, with a compute budget of 0.084 GFLOPs per clip that suggests real-time operation on mobile SoCs (exact throughput is device-dependent) and outperforming far heavier 3D CNN and transformer baselines.

Three design insights emerge from our study. First, re-parameterising temporal context into gated channel shifts can replace bulky 3D convolutions without loss of fidelity. Second, ultra-light channel attention is sufficient to recover much of the accuracy normally sacrificed when compressing backbones for edge hardware. Third, a carefully balanced sparse-frame strategy preserves recognition of fine-grained motions while keeping latency low.

Several limitations temper these findings. All experiments were confined to the 20BN-Jester dataset, leaving cross-dataset generalisation to outdoor lighting, cluttered backgrounds, and non-egocentric viewpoints untested. Inference speed was evaluated on a desktop GPU and inferred from FLOP counts; direct profiling on XR-class chips such as the Snapdragon XR2 Gen 2 or Apple Vision silicon is needed for a complete energy-latency picture. Finally, our prototype recognises only single-hand gestures, whereas many MR applications demand dual-hand or full-body interaction.

Future work will therefore extend the recogniser to bimanual and composite gesture sequences using a lightweight temporal-attention layer, incorporate inertial and spatial-audio cues via adaptive fusion to enhance robustness under motion blur and extreme lighting, and release an open-source implementation alongside on-device benchmarks and user studies that measure latency, precision, and subjective workload in realistic MR manipulation tasks. Taken together, the MobileNetV3 + G-DiTSM + DepthECA architecture offers a practical foundation for intuitive, low-latency 3D interaction in the next generation of mixed-reality systems.

All code, trained weights, and on-device benchmark scripts will be provided upon request

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 6816–6826. https://doi.org/10.1109/ICCV48922.2021.00676

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding?. In *Proceedings of the International Conference on Machine Learning (ICML)*. 813–824.

[3] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6299–6308.

[4] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. 2020. Skeleton-Based Action Recognition with Shift Graph Convolutional Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Quentin De Smedt, Hazem Wannous, and Jean-Philippe Vandeborre. 2016. Skeleton-based dynamic hand gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1206–1214.

[6] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2625–2634.

[7] Christoph Feichtenhofer. 2020. X3D: Expanding Architectures for Efficient Video Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 203–213.

[8] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. 2017. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense . In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 5843–5851. https://doi.org/10.1109/ICCV.2017.622

[9] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. 2019. Searching for MobileNetV3. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 1314–1324.

[10] ZhiYu Jiang, Yi Zhang, and Shu Hu. 2023. ESTI: an action recognition network with enhanced spatio-temporal information. *International Journal of Machine Learning and Cybernetics* 14, 9 (2023), 3059–3070.

[11] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager. 2016. Temporal Convolutional Networks for Action Segmentation and Detection. arXiv:1611.05267 [cs.CV] https://arxiv.org/abs/1611.05267

[12] Ji Lin, Chuang Gan, and Song Han. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 7083–7093.

[13] TIAN Ming LIU Jie, WANG Yue. 2023. Dynamic Gesture Recognition Network Based on Multiscale Spatiotemporal Feature Fusion. *Journal of Electronics & Information Technology* 45, 220758 (2023), 2614. https://doi.org/10.11999/JEIT220758

[14] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. 2019. The Jester Dataset: A Large-Scale Video Dataset of Human Gestures. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 2874–2882. https://doi.org/10.1109/ICCVW.2019.00349

[15] Yang Min, Liangli Zhang, Xiujuan He, and Hong Liu. 2020. PointLSTM: A Multi-Point LSTM Network for 3D Hand Gesture Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5761–5770.

[16] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz. 2016. Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks. In *CVPR*. 4207–4215. https://doi.org/10.1109/CVPR.2016.455

[17] Natalia Neverova, Christian Wolf, Graham W. Taylor, and François Nebout. 2016. ModDrop: Adaptive Multi-Modal Gesture Recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. 1692–1706.

[18] Lionel Pigou, Aäron van den Oord, Sander Dieleman, Mieke Van Herreweghe, and Joni Dambre. 2016. Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video. arXiv:1506.01911 [cs.CV] https://arxiv.org/abs/1506.01911

[19] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In *ICCV*.

[20] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 12026–12035.

[21] Karen Simonyan and Andrew Zisserman. 2014. Two-stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems (NeurIPS)*. 568–576.

[22] Mingxing Tan, Ruoming Pang, and Quoc V. Le. 2020. EfficientDet: Scalable and Efficient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10781–10790. https://doi.org/10.1109/CVPR42600.2020.01080

[23] Liu Ting-Long. 2024. Short-Term Action Learning for Video Action Recognition. *IEEE Access* 12 (2024), 30867–30875.

[24] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 4489–4497.

[25] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri. 2017. Convnet architecture search for spatiotemporal feature learning. *arXiv preprint arXiv:1708.05038* (2017).

[26] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6450–6459.

[27] Jun Wan, Stan Z. Li, Yibing Zhao, Shuai Zhou, Isabelle Guyon, and Sergio Escalera. 2016. ChaLearn Looking at People RGB-D Isolated and Continuous Datasets for Gesture Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 761–769. https://doi.org/10.1109/CVPRW.2016.100

[28] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal Segment Networks: Towards Good Practices for

Deep Action Recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 20–36.

[29] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2019. Temporal Segment Networks for Action Recognition in Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 11 (2019), 2740–2755. https://doi.org/10.1109/TPAMI.2018.2868668

[30] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local Neural Networks. arXiv:1711.07971 [cs.CV] https://arxiv.org/abs/1711.07971

[31] Zhengwei Wang, Qi She, and Aljosa Smolic. 2021. ACTION-Net: Multipath Excitation for Action Recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13209–13218. https://doi.org/10.1109/CVPR46437.2021.01301

[32] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 7444–7452.

[33] Can Zhang, Yuexian Zou, Guang Chen, and Lei Gan. 2020. Pan: Towards fast action recognition via learning persistence of appearance. *arXiv preprint arXiv:2008.03462* (2020).

[34] Kai Zhang, Wen-Huang Cheng, Yi-Hsin Chen, et al. 2018. EgoGesture: A Large-Scale Egocentric Hand Gesture Dataset. In *CVPR*. 841–849. https://doi.org/10.1109/CVPR.2018.00092

[35] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. 2017. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. arXiv:1707.01083 [cs.CV] https://arxiv.org/abs/1707.01083

[36] Enmin Zhong, Carlos R. del Blanco, Daniel Berjón, Fernando Jaureguizar, and Narciso García. 2023. Real-Time Monocular Skeleton-Based Hand Gesture Recognition Using 3D-Jointsformer. *Sensors* 23, 16 (2023), 7066. https://doi.org/10.3390/s23167066