S2Gest: Split-Scan State Space Models for Dynamic Hand Gesture Recognition

Kefan Chen[1], Yong Gu[2]✉, Bo Li[1], Longjie Huang[3], and Jiajun Zhang[1]

[1] Modern Industry School of Virtual Reality, Jiangxi University of Finance and Economics, Nanchang, China
[2] School of Software and Internet of Things Engineering, Jiangxi University of Finance and Economics, Nanchang, China
[3] School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang, China

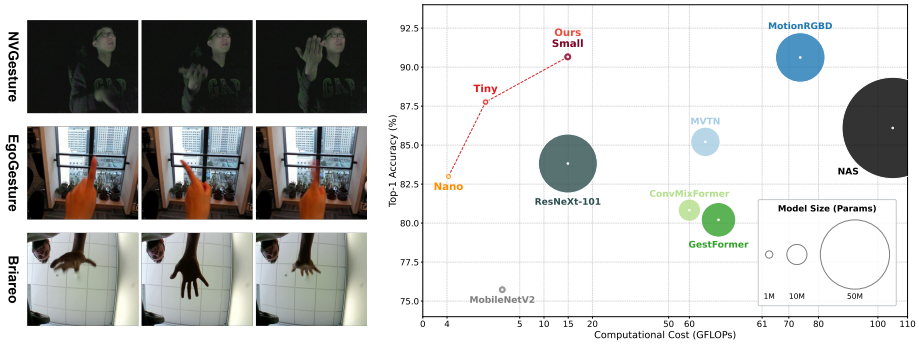
Abstract. Dynamic Hand Gesture Recognition is essential for natural human-computer interaction but faces challenges in continuous local execution on consumer-grade hardware due to strict computational and latency constraints. Existing solutions fail to reconcile the high Memory Access Cost (MAC) and dynamic footprint of Video Transformers with the limited temporal modeling capacity of lightweight CNNs. To address these limitations, we propose the Split-Scan Gesture Network (S2Gest), which rethinks temporal modeling by leveraging continuous state-space evolution to represent discrete sequential inputs. We introduce the Split-Scan Block (S2-Block), employing a novel tube-wise scanning strategy that partitions feature channels to process opposing temporal directions simultaneously. By structurally partitioning feature channels to capture complementary forward and backward dynamics, this design achieves global bidirectional modeling with linear-time complexity and without introducing additional parameters, effectively overcoming the causal unidirectionality inherent in standard State Space Models (SSMs). Evaluated across three benchmarks, the scalable S2Gest model family establishes a new state-of-the-art for lightweight architectures, matching the performance of large-scale baselines with a sub-3M parameter footprint while consistently exceeding 100 clips/s on consumer-grade GPUs. Code and weights are publicly available at `https://github.com/Chen-Ke-Fan/S2Gest`.

Keywords: Dynamic Gesture Recognition · State Space Models · Bidirectional Temporal Modeling · Lightweight Architecture

1 Introduction

Dynamic Hand Gesture Recognition (DHGR) has emerged as a critical intuitive interface for modern smart devices. However, always-on monitoring in personal spaces (*e.g.*, dimly lit vehicle interiors [25] or dynamic daily environments [34],

✉ Corresponding author. Email: iguyong@outlook.com



(a) *Challenging real-world scenarios.* **(b)** *Accuracy vs. efficiency on NVGesture (Depth).*

Fig. 1: Robustness and efficiency. (a) Typical local deployment challenges in consumer-grade scenarios include dynamic backgrounds and varying illumination. **(b)** Compared to large-scale baselines, S2Gest (Red) achieves state-of-the-art accuracy with significantly lower computational cost. Circle size represents parameter count.

Fig. 1a) makes local processing a strict privacy requirement rather than merely a latency benefit. This imposes stringent budgets on consumer-grade hardware across FLOPs, power, and runtime memory. Consequently, a fundamental conflict arises between the demand for robust recognition in complex environments and the severe constraints of continuous local deployment. Current architectures struggle to reconcile this conflict (Fig. 1b). Heavyweight models like 3D CNNs [1] and Video Transformers [8] incur prohibitive Memory Access Cost (MAC) and dynamic KV-cache expansion. Conversely, highly efficient lightweight CNNs [27] lack the temporal capacity to model complex kinematics.

To break this efficiency-accuracy bottleneck, we explore the potential of State Space Models (SSMs), specifically Mamba [11]. Beyond theoretical linear complexity, SSMs offer a structurally superior mechanism for processing continuous video streams on memory-constrained hardware. Unlike Transformers, which rely on explicit self-attention and a continuously expanding KV-cache to model temporal context, Mamba leverages a selective state-space formulation to compress historical information into a fixed-size hidden state. This enables it to act as a highly efficient recurrent memory container, effectively tracking long-range kinematic evolution without the prohibitive memory footprint. However, applying SSMs directly to DHGR presents two structural hurdles: (1) sparse depth inputs demand robust spatial preprocessing to extract structural semantics prior to temporal modeling, and (2) standard SSMs are inherently unidirectional, failing to capture the non-causal global context crucial for video understanding.

To bridge these gaps, we propose the Split-Scan Gesture Network (S2Gest). To address spatial sparsity efficiently, we introduce a Multi-Scale Spatial Block (MS-Block) adhering to a spatial-first embedding strategy, compressing redundancy early. Centrally, to resolve the unidirectional limitation of standard SSMs without escalating computational costs, we introduce the Split-Scan Block (S2-Block).

By structurally partitioning feature channels to scan forward and backward contexts simultaneously, the S2-Block explicitly decouples temporal directions. This achieves global bidirectional dynamics with linear-time complexity and without introducing additional parameters. Complementarily, an Adaptive Feature Modulation (AFM) module dynamically recalibrates the focal stream using global contextual priors to mitigate background noise.

Our contributions are summarized as follows:

- We propose S2Gest, an ultra-lightweight architecture tailored for continuous local DHGR, effectively reconciling the fundamental conflict between computational efficiency and complex temporal modeling.
- We design the S2-Block, integrating a Patch-Interleaved Tokenizer (PIT) and a novel Split-Scan mechanism. This enables bidirectional scanning without introducing additional parameters via structural channel partitioning.
- S2Gest achieves state-of-the-art (SOTA) lightweight performance on NVGesture [25], EgoGesture [34], and Briareo [23]. It matches heavyweight baselines while requiring significantly fewer parameters (*e.g.*, 2.59M *vs*. 48M), enabling real-time performance on consumer-grade GPUs.

2 Related Work

Spatiotemporal Modeling for DHGR. Early DHGR models predominantly relied on 3D CNNs [1, 15, 16] and multi-stream architectures [7, 14, 29, 35, 36] which, despite their accuracy, incur prohibitive computational costs due to dense convolutions. Subsequently, attention-based architectures [5, 6, 8, 9, 18, 30] achieved superior global contextual modeling. However, their quadratic $\mathcal{O}((THW)^2)$ complexity and dynamic KV-cache memory footprint render them impractical for local devices. This efficiency gap highlights a pressing need for architectures bridging Transformer-level global capacity with the hardware efficiency of compact 2D backbones.

Hardware-Aware Design & Edge Constraints. For local deployment, theoretical complexity often decouples from actual inference latency due to MAC [28, 32]. While lightweight image operators and re-parameterization techniques (*e.g.*, Xception [3], YOLOE [31]) reduce computational density, directly applying these static optimizations to video streams fails to exploit inherent temporal redundancy. Unlike these static approaches, our S2Gest navigates this constraint by employing a spatial-first embedding strategy, compressing spatial redundancy early to reserve the crucial computational budget for temporal modeling.

State Space Models for Video Understanding. State Space Models (SSMs) like Mamba [11] offer a constant-memory $\mathcal{O}(1)$ alternative to Transformers, making them intrinsically suited for kinematic tracking. To adapt SSMs for visual tasks, recent extensions [17, 21, 37] flatten 3D spatiotemporal tokens and employ multi-directional scanning (*e.g.*, cross-scan or bidirectional sweeps) to capture non-causal context. However, these methods perform parallel scanning over full,

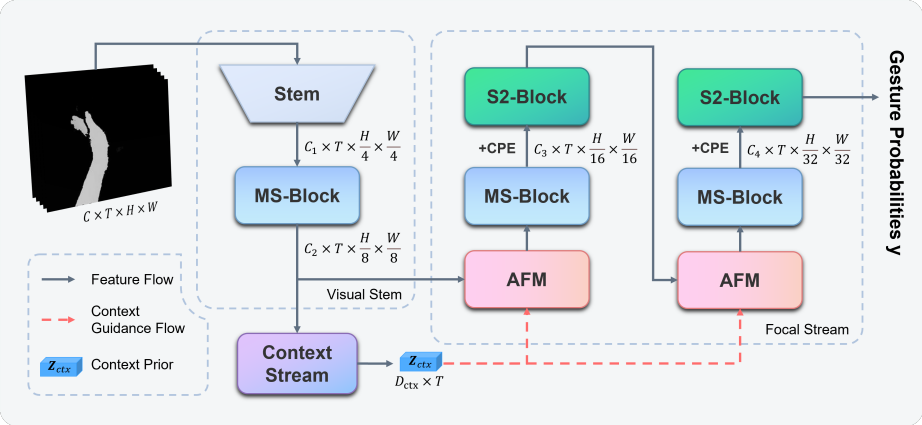


Fig. 2: Architecture of S2Gest. A shared Visual Stem feeds two streams: a lightweight Context Stream extracting global priors ($\mathbf{Z}_{\text{ctx}}$), and a high-resolution Focal Stream. Within the Focal Stream, AFM uses $\mathbf{Z}_{\text{ctx}}$ to dynamically modulate focal features (red dashed lines), which are then processed by bidirectional S2-Blocks for fine-grained trajectory modeling.

unpartitioned feature maps. This state-space replication effectively doubles the parameter count and memory footprint. To circumvent this exact limitation, S2Gest introduces a Split-Scan mechanism that achieves global non-causal context capture without introducing additional parameters, reconciling representation completeness with our sub-3M parameter goal.

3 Methodology

We introduce S2Gest, a unified framework designed to address the multi-objective constraints of continuous local deployment. We approach dynamic hand gesture recognition based on the premise that dynamic gestures are inherently continuous spatiotemporal kinematic evolutions, characterized by varying execution speeds and fluid transitions rather than isolated static poses.

Specifically, the system processes a sliding window of T consecutive frames during continuous deployment. As illustrated in Fig. 2, given a raw input video tensor $\mathbf{X}_{\text{in}} \in \mathbb{R}^{C\times T\times H\times W}$ (where C, H, W denote the channel, height, and spatial width, respectively), the pipeline operates in three sequential stages. A shared Visual Stem (Sec. 3.1) extracts localized structural embeddings while compressing spatial redundancy. The pipeline then bifurcates into our Context-Aware Dual-Stream architecture (Sec. 3.2): a lightweight Context Stream extracts global priors, which guide the high-resolution Focal Stream via Adaptive Feature Modulation (AFM). Before entering the deep temporal modeling phase, these modulated features are spatially anchored via Content-Adaptive Positional Encoding (CPE) [4]. Generated dynamically via a $3 \times 3 \times 3$ depthwise convolution, this standalone inductive bias establishes explicit 3D coordinates. Crucially, unlike

fixed absolute embeddings, this dynamic generation allows the entire pipeline to naturally extrapolate to arbitrary sequence lengths T during inference. Subsequently, within the Focal Stream, our cascaded S2-Blocks (Sec. 3.4) explicitly model bidirectional temporal dynamics across the sequence. A classification head then aggregates these spatiotemporal features to output the final gesture class probabilities $\mathbf{y} \in \mathbb{R}^K$ for K gesture categories.

3.1 Visual Stem: Spatial-First Embedding

Raw video frames exhibit substantial temporal redundancy early in the network, rendering dense 3D spatiotemporal processing computationally wasteful. Consequently, we adopt a "Spatial-First" strategy for the Visual Stem to compress redundancy while minimizing temporal computational overhead.

The stem integrates a hierarchical patch embedding layer with a Multi-Scale Spatial Block (MS-Block). To optimize for consumer-grade hardware, drawing inspiration from classic multi-scale representation learning, the MS-Block circumvents heavyweight 3D cubic convolutions ($k \times k \times k$) by adopting a highly efficient Inception-style topology. Formally, given an input feature tensor $\mathbf{X}$, the output is the concatenation of heterogeneous parallel branches:

$$\text{MS-Block}(\mathbf{X}) = \mathcal{H}_{\text{proj}}\left(\text{Concat}_{o\in\mathcal{O}}\left(\mathcal{B}_o(\mathbf{X})\right)\right), \tag{1}$$

where $\mathcal{O} = \{\text{Conv}_{1\times1\times1}, \text{Conv}^{(1)}_{1\times3\times3}, \text{Conv}^{(2)}_{1\times3\times3}, \text{MaxPool}_{3\times3\times3}\}$ denotes the operator set, and o indexes each operation branch. The dual $\text{Conv}_{1\times3\times3}$ branches enrich the receptive field variety, while the parameter-free $3 \times 3 \times 3$ MaxPool captures early local motion cues without the prohibitive MAC of 3D convolutions. $\mathcal{B}_o(\cdot)$ represents the corresponding transformation branch, and $\mathcal{H}_{\text{proj}}$ is a linear fusion projection. Crucially, this topology allows the spatial convolution and normalization layers to be fused (Conv-BN Fusion) during inference. This strategy maximizes throughput, reserving the computational budget for deep temporal modeling in the Focal Stream.

3.2 Context-Aware Dual-Stream Architecture

S2Gest implements a context-guided dual-stream pipeline where a lightweight context stream establishes global spatiotemporal priors to condition the high-resolution focal stream. This collaborative mechanism isolates global semantics from localized trajectories, optimizing computation on memory-constrained hardware.

The Context Stream branches from the Visual Stem features $\mathbf{X}_{\text{stem}}$, employing aggressive spatiotemporal downsampling to compress the representation. The resulting holistic context sequence $\mathbf{Z}_{\text{ctx}} \in \mathbb{R}^{D_{\text{ctx}}\times T}$ is obtained via spatial global average pooling:

$$\mathbf{Z}_{\text{ctx}} = \mathcal{P}_{\text{pool}}\left(\Phi_{\text{peri}}(\mathbf{X}_{\text{stem}})\right), \tag{2}$$

where $\Phi_{\text{peri}}(\cdot)$ denotes the spatiotemporal downsampling transformations of the Context Stream, and $\mathcal{P}_{\text{pool}}$ represents the spatial global pooling. Rather than

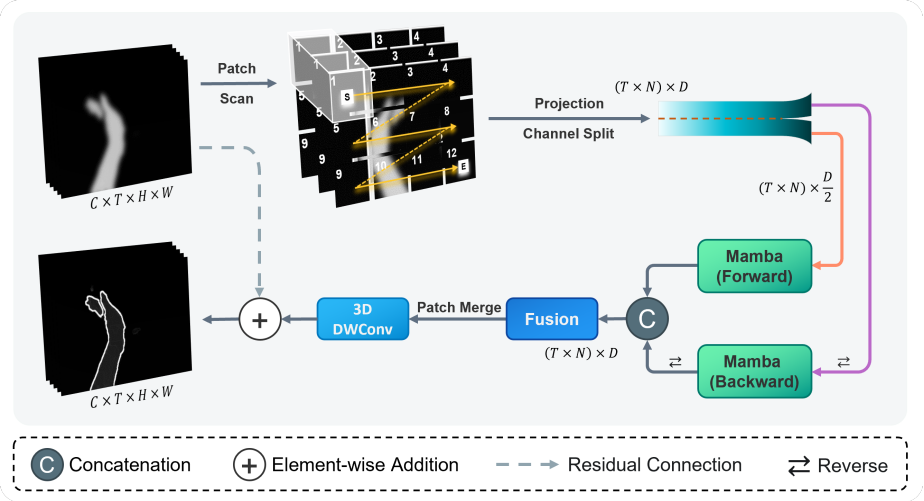


Fig. 3: Architecture of the S2-Block. Input features are directly serialized via the Patch-Interleaved Tokenizer (PIT, yellow scanning trajectory). The sequence is then explicitly decoupled along the channel dimension into two complementary subgroups. These are processed simultaneously by a parameter-neutral dual-stream SSM to capture bidirectional temporal dynamics. Finally, the streams are semantically fused and locally aggregated via a 3D depthwise convolution prior to residual addition.

processing full-resolution semantics which incurs prohibitive MAC, this ultra-lightweight design serves as an efficient prior extractor. By strictly constraining feature dimensions, it condenses global dynamics into a compact sequence with negligible overhead, preserving the primary computational budget for the fine-grained Focal Stream.

To instantiate conditional control over the main features, we employ Adaptive Feature Modulation (AFM) to dynamically recalibrate the Focal Stream. As illustrated in Fig. 2, this modulation acts as a critical bridge immediately after the Visual Stem, conditioning the features prior to deep temporal modeling. Let $\mathbf{F}_{\text{focal}}$ denote the feature tensor in the Focal Stream. Drawing inspiration from channel attention (*e.g.*, SE-Net [13]) and Feature-wise Linear Modulation (FiLM) [26], the context sequence $\mathbf{Z}_{\text{ctx}}$ is compressed into a global vector via Temporal Global Average Pooling (GAP). This vector subsequently generates channel-wise affine scaling and shifting parameters $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^C$ via learnable linear projections:

$$\mathbf{F}_{\text{mod}} = \text{AFM}(\mathbf{F}_{\text{focal}}, \mathbf{Z}_{\text{ctx}}) = \boldsymbol{\gamma} \odot \mathbf{F}_{\text{focal}} + \boldsymbol{\beta}, \tag{3}$$

where C is the channel dimension of $\mathbf{F}_{\text{focal}}$, and $\odot$ denotes element-wise multiplication with spatial and temporal broadcasting applied to $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. This modulation acts as a semantic filter, selectively reweighting channels to suppress background noise and enhance task-relevant kinematics before deep temporal scanning begins.

3.3 Continuous Dynamics via Selective SSMs

Rather than treating videos as discrete bags of frames, we model gesture execution as a continuous state-space evolution. To naturally align with the physical kinematic laws of hand movement, we formulate gesture recognition using continuous-time SSMs. The core state evolution is governed by a linear Ordinary Differential Equation (ODE) [2]:

$$\dot{\mathbf{h}}(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}x(t), \quad y(t) = \mathbf{C}\mathbf{h}(t), \tag{4}$$

where $x(t) \in \mathbb{R}$ is a 1-D continuous input signal, $\mathbf{h}(t) \in \mathbb{R}^N$ is the N-dimensional latent state, and $y(t) \in \mathbb{R}$ is the mapped output. The state transition matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ captures long-term memory, while $\mathbf{B} \in \mathbb{R}^{N\times 1}$ and $\mathbf{C} \in \mathbb{R}^{1\times N}$ act as projection parameters. Note that for multi-dimensional feature inputs in vision tasks, this formulation is applied independently across all channel dimensions.

In streaming inference scenarios, the system receives frames sequentially at fixed, discrete sampling intervals. However, the underlying physical motion and its semantic information rate are highly non-linear, featuring rapid swipes alongside static holds. To bridge the critical gap between evenly spaced discrete inputs and continuous physical dynamics, we employ Zero-Order Hold (ZOH) discretization with an input-dependent timescale $\Delta(\mathbf{x}_t)$, where $\mathbf{x}_t$ denotes the discrete sampled token at time step t. This data-dependent Δ allows the model to infer continuous dynamics from fixed-rate discrete frames. It acts as a dynamic gate: maintaining state memory during redundant static poses and driving fine-grained transitions during rapid motion. This formulation fundamentally accommodates the non-linear semantic rate of human kinematics better than rigid discrete tokenization.

3.4 S2-Block: Tube-wise Selective Motion Modeling

To instantiate the aforementioned continuous formulation, We propose the S2-Block to preserve local motion continuity under strict parameter constraints (Fig. 3). To achieve this, the block processes the feature tensor in two sequential phases: serializing the spatial map into Tube-Major trajectories, and routing them through a parameter-neutral bidirectional scanning engine.

Patch-Interleaved Tokenizer (PIT). To effectively align the spatiotemporal structure of video data with the 1D causal nature of SSMs, PIT enforces a Tube-Major scanning order. Standard "Frame-Major" serialization (*i.e.*, scanning all spatial tokens in frame t before $t+1$) fractures local motion continuity, forcing the SSM to memorize spatially disjoint tokens and causing severe *memory decay*. To resolve this, we permute the tensor axes to group tokens by spatial location first, then time. Formally, given the feature tensor $\mathbf{X} \in \mathbb{R}^{C\times T\times H\times W}$ (with channels C, frames T, and spatial resolution $H \times W$), we partition the spatial grid into $N = \frac{H}{P} \times \frac{W}{P}$ patches of size $P \times P$. The input is reshaped into spatiotemporal tubes:

$$\mathbf{X}_{\text{tube}} = \text{Permute}(\text{Patch}(\mathbf{X})) \in \mathbb{R}^{N\times T\times D_{\text{emb}}}, \tag{5}$$

where $D_{\text{emb}} = C \times P^2$ is the patch embedding dimension, and the sequence is organized as $\mathcal{S} = [\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_N]$. Each sub-sequence $\mathcal{T}_i = \{\mathbf{x}_{i,t} \mid t = 1 \ldots T\}$ tracks the complete temporal evolution of the i-th spatial patch. This serialization strategy directly aligns with the Eulerian perspective of motion analysis: unlike Frame-Major scanning which scatters temporal cues, PIT ensures that the SSM continuously observes dynamics within a fixed spatial aperture. While this introduces discontinuities between adjacent tubes, the SSM's selective gating mechanism ($\mathbf{\Delta}$) effectively resets its context at spatial boundaries, allowing the model to focus purely on robust *intra-tube* motion continuity.

Split-Scan: Parameter-Neutral Bidirectional Modeling. Standard causal SSMs struggle to capture retrospective dependencies essential for gesture semantic comprehension. While naive bidirectional augmentations (*e.g.*, Bi-Mamba, Video-Mamba) address this by feeding full-channel representations into dual branches, they incur a prohibitive 2× parameter and memory penalty. To resolve this context-efficiency dilemma, we propose a "Split-Transform-Merge" topology. Formally, given a serialized tube input $\mathbf{U} \in \mathbb{R}^{T \times D_{\text{emb}}}$, we explicitly decouple the channel space into two complementary subgroups, $\mathbf{U}_{\text{fwd}}$ and $\mathbf{U}_{\text{rev}}$, each with dimension $D_{\text{emb}}/2$. These are routed through a dual-stream system:

$$\mathbf{H}_{\text{fwd}} = \text{SSM}(\mathbf{U}_{\text{fwd}}), \tag{6a}$$

$$\mathbf{H}_{\text{rev}} = \text{Flip}(\text{SSM}(\text{Flip}(\mathbf{U}_{\text{rev}}))), \tag{6b}$$

$$\mathbf{Y} = \mathcal{F}_{\text{mix}}\left(\text{Norm}(\text{Concat}[\mathbf{H}_{\text{fwd}}, \mathbf{H}_{\text{rev}}])\right), \tag{6c}$$

where $\mathbf{H}_{\text{fwd}}$ and $\mathbf{H}_{\text{rev}}$ denote the directional hidden states. This design guarantees rigorous bidirectional modeling via explicit role assignment. The forward branch ($\mathbf{H}_{\text{fwd}}$) aligns with chronological execution, capturing the motion onset and unfolding trajectory. Conversely, the backward branch ($\mathbf{H}_{\text{rev}}$) processes the sequence in reverse, critical for capturing motion retraction and providing future global context to resolve temporal aliasing (*e.g.*, differentiating the preparatory phase of a complex gesture from its execution, as empirically verified in Sec. 4).

Crucially, structurally partitioning the channels explicitly halves the internal SSM transformation cost ($2 \times (D_{\text{emb}}/2)^2 = 0.5D_{\text{emb}}^2$) relative to a standard full-channel unidirectional baseline ($1.0D_{\text{emb}}^2$). The subsequent linear mixing stage ($\mathcal{F}_{\text{mix}}$) semantically integrates these complementary directional cues, restoring full-channel correlation while keeping the entire module strictly parameter-neutral. Finally, following this dual-stream scanning and fusion, the 1D sequence is reshaped back to its original 3D spatiotemporal layout (the "Patch Merge" stage in Fig. 3). To eliminate potential artificial spatial discontinuities introduced by the discrete tube boundaries, we apply a final 3D depthwise convolution. This acts as a local spatial smoothing filter prior to the residual connection, ensuring the absolute continuity of the refined motion representations.

4 Experiments

4.1 Experimental Settings

Datasets. We adhere to the official evaluation protocols for all datasets. NVGesture [25] (25 classes, 1,532 clips) features dimly lit vehicle interiors and varying illumination, serving to evaluate the model's robustness in visually constrained environments. EgoGesture [34] (83 classes, 24,161 clips) contains significant background clutter and ego-motion, providing a rigorous test for temporal robustness. Briareo [23] (12 classes, 1,440 clips) focuses on dynamic gestures for human-car interaction, introducing challenges related to unconventional viewpoints and complex hand articulations.

Implementation Details. Models are implemented in PyTorch and trained on NVIDIA A40 GPUs. We adopt a modality-aware initialization strategy: for RGB inputs, we utilize Jester [24] pre-training to facilitate semantic feature learning, whereas geometric modalities are trained entirely from scratch. We optimize the network using AdamW [22] with an initial learning rate of 5×10^{-5}, a multi-step decay schedule, and a weight decay of 0.05. The batch size is set to 4. We empirically observed that this small batch size serves as an implicit regularizer, improving generalization on high-dimensional video data. Input clips are uniformly sampled to a sequence length of $T = 40$ frames, with each frame resized to a spatial resolution of 192×256, ensuring sufficient temporal context and spatial fidelity. Standard spatial augmentations (random cropping, scaling, and rotation) are applied.

Evaluation Metrics. We report Top-1 Accuracy (%) to measure recognition performance. To thoroughly evaluate model efficiency, we assess computational complexity using the number of parameters (M) and FLOPs (G). Furthermore, to gauge practical deployment capabilities, we independently measure real-world latency (ms) and throughput (clips/s).

4.2 Comparison with State-of-the-Art

Tab. 1 compares S2Gest with state-of-the-art approaches across RGB and Depth modalities. To ensure rigorous evaluation, we report the Top-1 accuracies and computational complexities (Params and GFLOPs) directly from their original publications. For models where complexity metrics are unavailable (marked with †), we independently compute them using a standard input sequence length of $T = 40$. Furthermore, to maintain a fair comparison, we refrain from evaluating unofficial reproductions of complex baselines, which may exhibit sub-optimal performance.

While baselines may employ varying optimal spatial resolutions and clip lengths, their settings remain within a comparable temporal scale, ensuring meaningful comparability. Notably, unlike several baselines that rely on extensive multi-modal pre-training, our Depth models achieve competitive results trained entirely from scratch, demonstrating superior data efficiency. For completeness, we also include a standard MobileNetV2 as a lightweight reference.

Table 1: Comparison with SOTA methods for dynamic hand gesture recognition. We evaluate the Top-1 accuracy (%) and computational complexity on three standard benchmarks. The results demonstrate that our proposed S2Gest family provides a highly lightweight alternative without sacrificing recognition accuracy.

Method	Params	FLOPs	NVGesture		EgoGesture		Briareo	
	(M)	(G)	RGB	Depth	RGB	Depth	RGB	Depth
NAS [33]	127.10	120.20	83.61	86.10	93.31	94.13	–	–
ResNeXt-101† [15]	47.53	14.93	78.63	83.82	93.75	94.03	–	–
MotionRGBD† [36]	38.14	73.85	**89.58**	90.62	–	–	–	–
GestFormer [8]	24.08	60.40	75.41	80.21	–	–	94.44	96.18
MDRA [20]	23.90	45.30	83.87	86.60	93.96	94.20	97.84	96.92
MVTN [9]	19.55	60.22	77.50	85.21	–	–	97.69	97.92
ConvMixFormer [10]	13.57	59.98	76.04	80.83	–	–	**98.26**	97.22
DSTCNet [19]	9.60	42.56	82.50	88.33	93.70	94.26	97.81	98.10
MobileNetV2 1.0* [27]	2.38	4.76	74.07	75.73	93.49	94.18	93.52	91.20
S2Gest (Ours)								
– Small	2.59	14.90	86.10	**90.66**	**96.47**	**97.35**	97.69	98.15
– Tiny	1.17	4.53	85.06	87.76	96.27	97.27	97.69	**98.61**
– Nano	0.68	4.02	85.48	82.99	95.94	97.07	98.15	97.69

∗: All metrics are fully reproduced by us using the official code. †: Complexity is re-evaluated by us; accuracy is cited.
Bold (red) and underlined (blue) values indicate the best and second-best performance, respectively.

Competitive Performance on NVGesture. On NVGesture, S2Gest-Small achieves a highly competitive 90.66% Depth accuracy, matching the heavyweight MotionRGBD [36] (90.62%) with a ∼15× reduction in parameters (2.59M vs. 38.14M). Furthermore, our ultra-lightweight Nano variant outperforms MobileNetV2 [27] by 7.26% in Depth using only 29% of its parameters, proving highly data-efficient on complex tasks.

Superior Robustness on EgoGesture and Briareo. On EgoGesture and Briareo, S2Gest-Small establishes new SOTA performance. It achieves 96.47% (RGB) and 97.35% (Depth) on EgoGesture, outperforming NAS [33]. On Briareo, S2Gest-Tiny reaches a peak 98.61% Depth accuracy, surpassing ConvMixFormer [10]. Remarkably, the Nano variant (0.68M Params, 4.02 GFLOPs) remains robust (97.07% EgoGesture Depth; 97.69% Briareo Depth), making it ideal for resource-constrained deployment.

Insights on Modality Discrepancy and Model Capacity. Our results reveal two critical insights regarding modality sensitivity and architectural scaling. First, we observe a significant performance gap between modalities on NVGesture, where S2Gest-Small RGB trails Depth by 4.56% (86.10% vs. 90.66%). This discrepancy is fundamentally rooted in the dataset's constrained environments; RGB features are highly sensitive to dimly lit vehicle interiors and varying illumination, whereas Depth maps provide illumination-invariant geometric cues, yielding far greater robustness.

Second, our evaluation highlights the counter-intuitive advantage of ultra-lightweight models in noisy environments. On datasets characterized by significant

Table 2: Component-wise ablation study on NVGesture (Depth). We validate the effectiveness of our proposed components: MS-Block (Spatial), S2-Block (Temporal), AFM (Guidance), and Split (Scanning).

#	Spatial	Temporal	Guidance	Scanning	Params (M)	Acc (%)
0	MS-Block	S2-Block	AFM	Split-Scan	2.59	**90.66**
1	3D ResNet-18* [12]	–	–	–	2.66	83.82
2	–	Transformer [30]	–	N/A	2.79	86.51
3	–	–	None	–	2.24	88.80
4	–	–	Add	–	2.53	87.76
5	–	–	Concat	–	2.52	89.21
6	–	–	–	Uni-Direct.	2.80	89.42
7	–	–	–	Full Bi-Direct.	3.26	90.04

∗: Channels reduced to < 3M parameters. Row 0: Our S2Gest baseline. "–": Identical to Row 0.

background clutter (*e.g.*, EgoGesture), extreme down-scaling does not lead to severe underfitting. Instead, the restricted parameter space of S2Gest-Nano (0.68M) acts as an implicit regularizer. By preventing the network from overfitting to irrelevant background pixels, the capacity bottleneck forces the model to encode only the most essential temporal motion semantics. Consequently, Nano maintains a near-ceiling performance ($> 97\%$), occasionally matching or eclipsing its larger counterparts (*e.g.*, Nano RGB achieves 98.15% vs. 97.69% for both Tiny and Small on Briareo), proving highly advantageous for complex, noisy real-world deployments.

4.3 Ablation Studies

We conduct comprehensive ablation studies on the NVGesture dataset (Depth modality) to validate the effectiveness of each proposed component. As summarized in Tab. 2, we systematically isolate the contributions of the Spatial Encoder, the core Temporal Modeling Module, the Contextual Guidance, and the Split-Scan Mechanism. All models are trained under identical settings.

Effectiveness of Spatial Encoder. To validate the representational capacity of our spatial stem, we substitute the proposed MS-Block with a 3D adaptation of ResNet-18 [12] (Row 1), deliberately scaling its channels to fit the sub-3M parameter constraint. This substitution results in a significant performance degradation of 6.84%. The substantial drop indicates that our multi-scale receptive fields are highly effective at extracting fine-grained spatial kinematics under strictly bounded computational budgets.

Mamba vs. Transformer (Temporal). As shown in Row 2, replacing the S2-Block with a standard Transformer Encoder [30] incurs a 4.15% performance drop. This demonstrates that the continuous state-space modeling of Mamba aligns more naturally with the continuous spatiotemporal dynamics of hand gestures than discrete attention mechanisms.

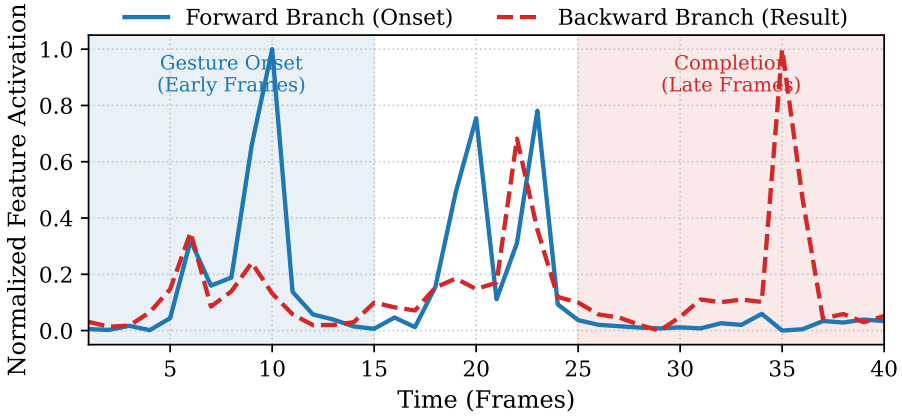


Fig. 4: Interpretability of Split-Scan Activations. Asymmetric branch peaks demonstrate decoupled, non-redundant feature extraction. The forward branch acts as a causal trigger at the gesture's onset, while the backward branch serves as an anti-causal validator at completion, effectively preventing feature collapse.

Necessity of AFM Guidance. Rows 3-5 provide crucial insights into context integration. Interestingly, naive element-wise addition (Row 4, 87.76%) performs worse than employing no guidance (Row 3, 88.80%). This suggests that the unconditional injection of global context as a bias term introduces representational noise, disrupting local feature learning. In contrast, our AFM (Row 0, 90.66%) employs adaptive modulation to dynamically recalibrate features, effectively leveraging global priors to boost performance.

Optimality of Split-Scan Strategy. A uni-directional scan (Row 6) yields suboptimal results (89.42%) due to the lack of retrospective context. Notably, the full bi-directional strategy (Row 7) increases the model size to 3.26M yet slightly underperforms our baseline (90.04%), likely due to channel redundancy and overlapping feature representations. Conversely, our Split strategy (Row 0) encodes complementary bidirectional dependencies within disjoint channel groups. By partitioning the C channels into two $C/2$ streams, the quadratic scaling of the linear projection matrices is significantly constrained. This acts as a structural regularizer, simultaneously reducing overall parameters (2.59M vs. 2.80M) and enhancing feature generalization.

Interpretability of Split-Scan. To further validate the mechanistic design of this structural regularizer, Fig. 4 visualizes the branch activations during inference. Dynamic gestures are fundamentally defined by temporal boundaries. As the visualizations reveal, the forward branch typically acts as a *causal trigger*, sharply peaking at the gesture's onset. Conversely, the backward branch functions as an *anti-causal validator*, peaking at the gesture's completion or retraction. This highly asymmetric firing empirically demonstrates that our explicit channel partitioning

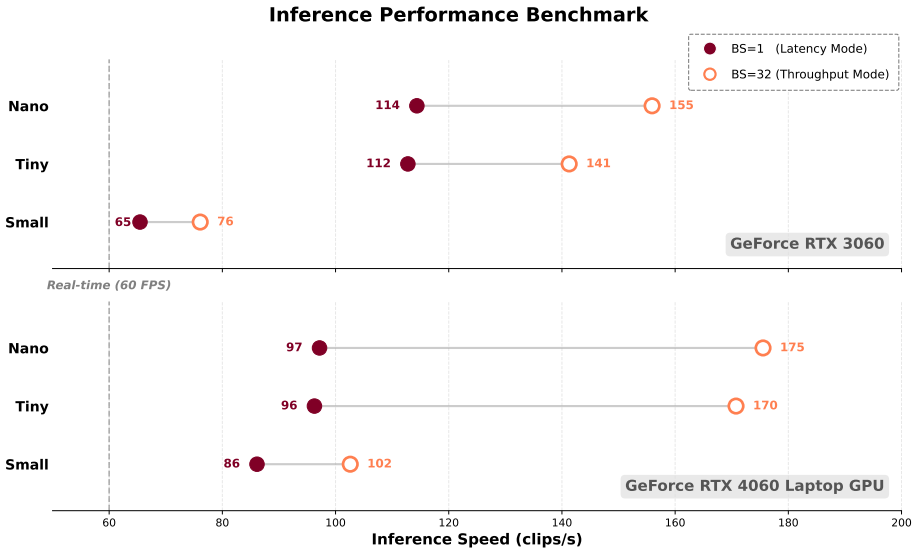


Fig. 5: Real-world Inference Benchmark. We report the throughput in clips/s on a desktop RTX 3060 (Top) and an RTX 4060 Laptop (Bottom). The latency mode (BS=1) simulates real-time interaction, while the throughput mode (BS=32) simulates batched processing. Notably, all S2Gest variants comfortably surpass the real-time threshold (dashed line) even on standard consumer-grade GPUs.

prevents feature collapse, a common redundancy flaw in naive Bi-Mamba designs, and successfully extracts decoupled, complementary semantics.

4.4 Efficiency and Inference Analysis

Beyond theoretical complexity (parameters and FLOPs), explicitly evaluating the trade-off between recognition performance and real-world latency is crucial for sustained local deployment. We benchmark our S2Gest family on typical consumer-grade GPUs (Desktop NVIDIA RTX 3060 and Laptop RTX 4060) to evaluate its *local-ready* capability for AI PCs and smart cockpits. Fig. 5 visualizes the performance in both latency mode (batch size = 1) and throughput mode (batch size = 32).

Interactive Inference Speed (BS=1). For interactive applications where responsiveness is paramount, our models exhibit rapid execution speed. On the desktop RTX 3060, our highest-accuracy S2Gest-Small achieves 65.45 clips/s (∼15.28 ms per clip), safely exceeding the standard real-time requirement of 30 clips/s. For ultra-low latency scenarios, S2Gest-Nano reaches a peak of 114.37 clips/s (∼8.74 ms), making it ideal for instant gesture feedback systems such as virtual reality (VR) and augmented reality (AR) interfaces.

High-Throughput Processing (BS=32). When prioritization shifts to processing throughput (*e.g.*, offline log analysis), our architecture scales highly efficiently. Operating under compute-bound conditions at BS=32, the Tiny and Nano variants achieve remarkable throughputs of 170.76 and 175.53 clips/s on the Laptop RTX 4060. Evaluated under identical settings, a traditional lightweight baseline like MobileNetV2 is only marginally faster (181 clips/s) yet requires $> 3.5\times$ the parameters of S2Gest-Nano and yields consistently lower accuracy. Furthermore, unlike Transformer-based architectures that suffer from severe dynamic memory overhead at high throughput, our sustained performance empirically demonstrates that S2Gest effectively circumvents the memory-wall bottleneck and KV-cache expansion typical of attention-based mechanisms.

Insights on Hardware Dynamics: Memory vs. Compute. Interestingly, a cross-device analysis in Fig. 5 reveals a subtle hardware-driven throughput inversion, highlighting the operational dynamics of our architecture. At BS=1, where ultra-lightweight models are heavily memory-bound, the laptop RTX 4060 trails the desktop RTX 3060 on the Nano and Tiny variants due to its narrower memory bandwidth (256 vs. 360 GB/s). Conversely, at BS=32, the workload becomes predominantly compute-bound. Here, the RTX 4060's superior compute density and performance-per-watt enable it to universally outperform the desktop RTX 3060 across all variants, despite a strictly restricted power limit ($\sim$100W vs. 170W).

Extreme Edge Viability. Beyond GPUs, our lightweight design is a critical enabler for real-world CPU-bound edge systems. On highly constrained low-power CPUs where sustained compute often falls below 15 GFLOPs/s, heavy baselines (> 20 GFLOPs) suffer severe computational bottlenecks, precluding real-time human-computer interaction. In stark contrast, the minimal computational footprint of S2Gest-Nano (4.02 GFLOPs) theoretically unlocks the viability of deploying deep spatiotemporal modeling directly onto resource-deprived edge devices, drastically lowering the hardware barrier for interactive applications in future deployments.

4.5 Discussion and Limitations

While S2Gest demonstrates strong efficiency and robustness, it exhibits specific failure modes under extreme conditions. Spatially, the aggressive early downsampling within our Visual Stem, designed to satisfy strict computational budgets, can occasionally obscure fine-grained articulated finger boundaries. Consequently, under severe motion blur, the model struggles to disambiguate highly similar topological configurations (*e.g.*, misclassifying "swiping with two fingers" as "three fingers"). Temporally, our reliance on ZOH discretization operates under the assumption of relatively stable execution speeds. During highly erratic or rapid cyclical gestures, severe temporal aliasing can disrupt the selective gating mechanism of the bidirectional SSM, occasionally leading to premature trajectory termination. Addressing these bottlenecks through adaptive token preservation and dynamic ODE solvers presents a compelling direction for future work.

5 Conclusion

In this paper, we presented S2Gest, an ultra-lightweight spatiotemporal architecture designed to resolve the conflict between high-performance gesture recognition and the strict constraints of continuous local deployment. By rethinking temporal modeling through the lens of continuous state-space evolution, we introduced the S2-Block equipped with a novel Split-Scan strategy. This mechanism, grounded in decoupled temporal factorization, effectively overcomes the unidirectionality of standard SSMs, achieving global bidirectional modeling with linear-time complexity and zero parameter overhead. Extensive experiments on NVGesture, EgoGesture, and Briareo demonstrate that S2Gest establishes a new state-of-the-art among lightweight models. It delivers robust accuracy comparable to computationally heavy 3D CNNs while maintaining exceptional real-time throughput across diverse hardware landscapes, ranging from consumer-grade GPUs to theoretical viability on highly constrained edge CPUs.

Beyond gesture recognition, we believe the Split-Scan mechanism and tube-wise modeling provide a versatile building block for efficient video understanding. Future research will explore extending this architecture with noise-robust spatial stems and multi-modal fusion, opening new avenues for high-performance, real-time perception on resource-constrained local platforms.

Acknowledgements

This work was supported by the Postgraduate Innovation Special Fund Project of Jiangxi Province under Grant YC2025-S475, and the 20th Student Scientific Research Project of Jiangxi University of Finance and Economics under Grant 20251211191926034.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4733. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.502
2. Chen, T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.: Neural ordinary differential equations. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 6572–6583 (2018), https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html
3. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1800–1807. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.195
4. Chu, X., Tian, Z., Zhang, B., Wang, X., Shen, C.: Conditional positional encodings for vision transformers. In: International Conference on Learning Representations (ICLR). OpenReview.net (2023), https://openreview.net/forum?id=3KWnuT-R1bh

5. D'Eusanio, A., Simoni, A., Pini, S., Borghi, G., Vezzani, R., Cucchiara, R.: A transformer-based network for dynamic hand gesture recognition. In: International Conference on 3D Vision (3DV). pp. 623–632. IEEE (2020). https://doi.org/10.1109/3DV50981.2020.00072
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR). OpenReview.net (2021), https://openreview.net/forum?id=YicbFdNTTy
7. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6201–6210. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00630
8. Garg, M., Ghosh, D., Pradhan, P.M.: Gestformer: Multiscale wavelet pooling transformer network for dynamic hand gesture recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2473–2483. IEEE (2024). https://doi.org/10.1109/CVPRW63382.2024.00254
9. Garg, M., Ghosh, D., Pradhan, P.M.: MVTN: A multiscale video transformer network for hand gesture recognition. In: European Conference on Computer Vision Workshops (ECCVW). Lecture Notes in Computer Science, vol. 15644, pp. 15–33. Springer (2024). https://doi.org/10.1007/978-3-031-92089-9_2
10. Garg, M., Ghosh, D., Pradhan, P.M.: Convmixformer- A resource-efficient convolution mixer for transformer-based dynamic hand gesture recognition. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6156–6166. IEEE (2025). https://doi.org/10.1109/WACV61041.2025.00600
11. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. CoRR **abs/2312.00752** (2023). https://doi.org/10.48550/ARXIV.2312.00752
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.90
13. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7132–7141. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00745
14. Joze, H.R.V., Shaban, A., Iuzzolino, M.L., Koishida, K.: MMTM: multimodal transfer module for CNN fusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13286–13296. Computer Vision Foundation / IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.01330
15. Köpüklü, O., Gunduz, A., Kose, N., Rigoll, G.: Real-time hand gesture detection and classification using convolutional neural networks. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG). pp. 1–8. IEEE (2019). https://doi.org/10.1109/FG.2019.8756576
16. Köpüklü, O., Kose, N., Gunduz, A., Rigoll, G.: Resource efficient 3d convolutional neural networks. In: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 1910–1919. IEEE (2019). https://doi.org/10.1109/ICCVW.2019.00240
17. Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: Videomamba: State space model for efficient video understanding. In: European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 15084, pp. 237–255. Springer (2024). https://doi.org/10.1007/978-3-031-73347-5_14

18. 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. **14**(1), 246–252 (2022). https://doi.org/10.1109/TCDS.2020.3048883
19. Liao, G., Chen, K., Huang, L., Gu, Y., Li, B.: A lightweight multi-variable spatio-temporal convolutional framework for dynamic gesture recognition. In: International Conference on 3D Visio (3DV). pp. 936–945. IEEE (2026). https://doi.org/10.1109/3DV69130.2026.00094
20. Liao, G., Huang, L., Chen, K., Gu, Y., Zhu, T.: MDRA: A motion-guided dual-stream recurrent attention framework for dynamic hand gesture recognition. ACM Trans. Multim. Comput. Commun. Appl. **22**(5), 149:1–149:19 (2026). https://doi.org/10.1145/3797044
21. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. In: Advances in Neural Information Processing Systems (NeurIPS) (2024), http://papers.nips.cc/paper_files/paper/2024/hash/baa2da9ae4bfed26520bb61d259a3653-Abstract-Conference.html
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR). OpenReview.net (2019), https://openreview.net/forum?id=Bkg6RiCqY7
23. Manganaro, F., Pini, S., Borghi, G., Vezzani, R., Cucchiara, R.: Hand gestures for the human-car interaction: The briareo dataset. In: International Conference on Image Analysis and Processing (ICIAP). Lecture Notes in Computer Science, vol. 11752, pp. 560–571. Springer (2019). https://doi.org/10.1007/978-3-030-30645-8_51
24. Materzynska, J., Berger, G., Bax, I., Memisevic, R.: The jester dataset: A large-scale video dataset of human gestures. In: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 2874–2882. IEEE (2019). https://doi.org/10.1109/ICCVW.2019.00349
25. Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., Kautz, J.: Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4207–4215. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.456
26. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.: Film: Visual reasoning with a general conditioning layer. In: AAAI Conference on Artificial Intelligence (AAAI). pp. 3942–3951. AAAI Press (2018). https://doi.org/10.1609/AAAI.V32I1.11671
27. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4510–4520. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00474
28. Sinha, N., Rostami, P., Shabayek, A.E.R., Kacem, A., Aouada, D.: Multi-objective hardware aware neural architecture search using hardware cost diversity. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 8032–8039. IEEE (2024). https://doi.org/10.1109/CVPRW63382.2024.00802
29. Tang, S., Zhu, M., Peng, Y., Zheng, L.: Dste-net: Dual-scale spatial-temporal excitation network for dynamic gesture recognition. IEEE Robotics Autom. Lett. **10**(12), 12844–12851 (2025). https://doi.org/10.1109/LRA.2025.3625493
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 5998–6008 (2017), https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

31. Wang, A., Liu, L., Chen, H., Lin, Z., Han, J., Ding, G.: YOLOE: real-time seeing anything. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 24591–24602. IEEE (2025). https://doi.org/10.1109/ICCV51701.2025.02280
32. Xiao, J., Zhang, C., Gong, Y., Yin, M., Sui, Y., Xiang, L., Tao, D., Yuan, B.: HALOC: hardware-aware automatic low-rank compression for compact neural networks. In: AAAI Conference on Artificial Intelligence (AAAI). pp. 10464–10472. AAAI Press (2023). https://doi.org/10.1609/AAAI.V37I9.26244
33. Yu, Z., Zhou, B., Wan, J., Wang, P., Chen, H., Liu, X., Li, S.Z., Zhao, G.: Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. IEEE Trans. Image Process. **30**, 5626–5640 (2021). https://doi.org/10.1109/TIP.2021.3087348
34. Zhang, Y., Cao, C., Cheng, J., Lu, H.: Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. IEEE Trans. Multim. **20**(5), 1038–1050 (2018). https://doi.org/10.1109/TMM.2018.2808769
35. Zhou, B., Li, Y., Wan, J.: Regional attention with architecture-rebuilt 3d network for RGB-D gesture recognition. In: AAAI Conference on Artificial Intelligence (AAAI). pp. 3563–3571. AAAI Press (2021). https://doi.org/10.1609/AAAI.V35I4.16471
36. Zhou, B., Wang, P., Wan, J., Liang, Y., Wang, F., Zhang, D., Lei, Z., Li, H., Jin, R.: Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20122–20131. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01952
37. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. In: International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 235, pp. 62429–62442. PMLR / OpenReview.net (2024), https://proceedings.mlr.press/v235/zhu24f.html