A Lightweight Multi-Variable Spatio-Temporal Convolutional Framework for Dynamic Gesture Recognition

Guoqiong Liao[1,2] Kefan Chen[1†] Longjie Huang[3] Yong Gu[4] Bo Li[1]

[1]Modern Industry School of Virtual Reality, Jiangxi University of Finance and Economics, China
[2]Jiangxi Tourism & Commerce Vocational College, China
[3]School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, China
[4]School of Software and Internet of Things Engineering, Jiangxi University of Finance and Economics, China

Abstract

Transformer-based hybrid architectures have achieved remarkable performance in dynamic hand gesture recognition. However, their high computational overhead and model size limit deployment in resource-limited environments. Motivated by this limitation, we propose the Decoupled Spatio-Temporal Convolutional Network (DSTCNet), a lightweight, pure convolutional framework trained end-to-end delivering high accuracy with a fraction of the complexity. DSTCNet integrates two components: (1) an efficient pseudo-3D spatial backbone, the Pseudo-3D Gated Attentional Fusion Network (P3D-GAFNet), enhancing spatial feature extraction via positional prior injection, and (2) a temporal modeling network, the Multi-Variable Decomposition Temporal Convolutional Network (MVD-TCN), leveraging multi-variable feature decomposition with modern convolutional blocks to capture long-range temporal dependencies without the cost of self-attention. With only 9.6M parameters, DSTCNet matches or surpasses the accuracy of substantially larger models on several challenging benchmarks, while offering high computational efficiency, lower memory usage, and reduced energy consumption—making it a practical solution for deployment on edge devices. Our results demonstrate that modernized pure convolutional architectures can serve as a robust and efficient alternative to hybrid designs, offering valuable insights for the broader field of video understanding.

1. Introduction

Dynamic hand gesture recognition is a cornerstone technology in Human-Computer Interaction, with key applications in virtual reality, intelligent cockpits, and sign language translation. The widespread adoption of low-cost RGB-D sensors has significantly advanced the field. However, in practice, different scenarios have distinct requirements. For instance, typically resource-constrained edge devices need lightweight and efficient models, whereas domains such as healthcare and rehabilitation demand precise and robust recognition. These diverse requirements pose a fundamental challenge: how to design a single framework that can simultaneously offer efficiency, robustness, and adaptability to different deployment environments.

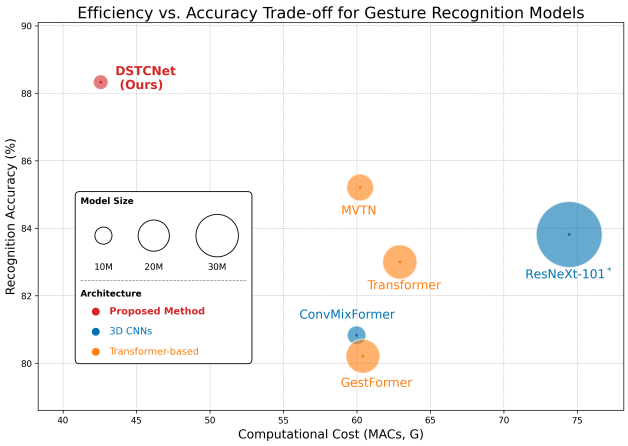

Figure 1. Comparison of DSTCNet (red) against various prevalent methods. Results with * are reproduced by the authors.

The pursuit of higher accuracy has recently been dominated by Transformer-based and hybrid architectures [10, 11, 17]. However, this paradigm, while powerful, is fundamentally misaligned with the constraints of practical deployment, suffering from a dual bottleneck of computational cost and architectural bloat. At its core, the quadratic scaling of self-attention imposes a significant computational barrier, limiting throughput and increasing energy consumption on the long sequences typical of gesture recognition. Furthermore, their design philosophy encourages

†Corresponding author: chen_kefan@outlook.com

deep stacks of parameter-heavy blocks, resulting in oversized models that create a severe deployment bottleneck for edge devices with limited memory and storage. These combined issues create a widening gap between benchmark performance and real-world deployability, forcing a severe trade-off between accuracy and practicality. This challenge is visually summarized in Figure 1.

Concurrently, recent research in time-series analysis [18] has demonstrated that modernized purely convolutional networks can achieve strong long-term temporal modeling capabilities without attention mechanisms. This line of work suggests that convolution-based approaches, long regarded as efficient yet less expressive than Transformers, may in fact offer a competitive alternative when properly modernized. Building upon this perspective, a natural question arises: can dynamic hand gesture recognition completely break free from Transformer architectures, achieving significant efficiency gains while simultaneously maintaining or even enhancing performance? Addressing this question not only provides a new research direction but also carries profound implications for the practical scalability and sustainability of gesture recognition technologies.

Motivated by the limitations of Transformer-based designs and inspired by the efficiency of modern convolutional approaches, we propose the Decoupled Spatio-Temporal Convolutional Network (DSTCNet), a fully convolutional framework for dynamic hand gesture recognition. The framework employs a decoupled architecture composed of two specialized modules. First, the Pseudo-3D Gated Attentional Fusion Network (P3D-GAFNet), a pseudo-3D spatial backbone, utilizes our proposed Gated Attentional Fusion (GAF) module to incorporate positional priors, enabling context-aware spatial feature extraction. Second, the Multi-Variable Decomposition Temporal Convolutional Network (MVD-TCN), a temporal modeling network, leverages multi-variable feature decomposition alongside modern large-kernel convolutions to capture long-range temporal dependencies. Extensive evaluations on multiple RGB-D benchmarks demonstrate that DSTC-Net achieves performance competitive with or superior to existing architectures while significantly reducing the parameter count and computational overhead.

Our main contributions are as follows:

1. We propose a lightweight, end-to-end purely convolutional framework for gesture recognition. With only 9.6M parameters, it achieves competitive performance with superior computational efficiency.
2. We design P3D-GAFNet, which incorporates a GAF module to inject positional priors for precise spatial feature extraction.
3. We introduce MVD-TCN, a temporal modeling network using Multi-Variable Feature Decomposition to capture long-range dependencies and complex dynamics.

2. Related Work

Dynamic hand gesture recognition has progressed from handcrafted features and classical models (e.g., SVMs, HMMs) to deep learning methods capturing temporal dynamics with RNNs and LSTMs, and more recently to 3D-CNNs for powerful joint spatio-temporal modeling. These milestones paved the way for modern architectures, motivating the exploration of more efficient and effective modeling strategies.

2.1. Transformers and Hybrid Models

The Transformer architecture [24], with its self-attention mechanism, excels at capturing long-range dependencies in gesture recognition [7], but its quadratic computational complexity remains a bottleneck for high-resolution or long videos. To address this, a popular research direction involves creating hybrid architectures that augment Transformers with convolutions. Some methods, like MsMHA-VTH [8], incorporate multi-scale features into the attention mechanism, while others, like GestFormer [9], use pre-processing modules such as wavelet pooling. More direct approaches, such as ConvMixFormer [11], replace expensive self-attention with lightweight convolutional mixers. These designs successfully inject beneficial convolutional inductive biases, reducing parameters and computation while maintaining strong performance. However, they remain incremental improvements on the Transformer backbone, inheriting its core architectural limitations and motivating the exploration of radically different, pure convolutional alternatives.

2.2. Pure Convolutional Temporal Modeling

Temporal Convolutional Networks (TCNs) [2] were early pure convolutional models for sequence modeling, employing causal and dilated convolutions. While offering parallelism and stability, their limited effective receptive field (ERF) restricted long-range dependency modeling, often requiring deep stacking that diminished their efficiency advantages over Transformers.

Inspired by Transformers, modern convolutional networks have since evolved. In computer vision, ConvNeXt [16] adapted Transformer design principles, RepLKNet [6] demonstrated that large kernels overcome the ERF bottleneck with structural re-parameterization, and MetaFormer [28] argued that the general architecture, not just self-attention, drives performance. Although these works renewed interest in CNNs, they mainly focused on static tasks. This philosophy was recently extended to time-series analysis with ModernTCN [18], which integrates large kernels and decoupling strategies to achieve state-of-the-art results. However, a research gap persists in adapting these methods from low-dimensional signals to the complex,

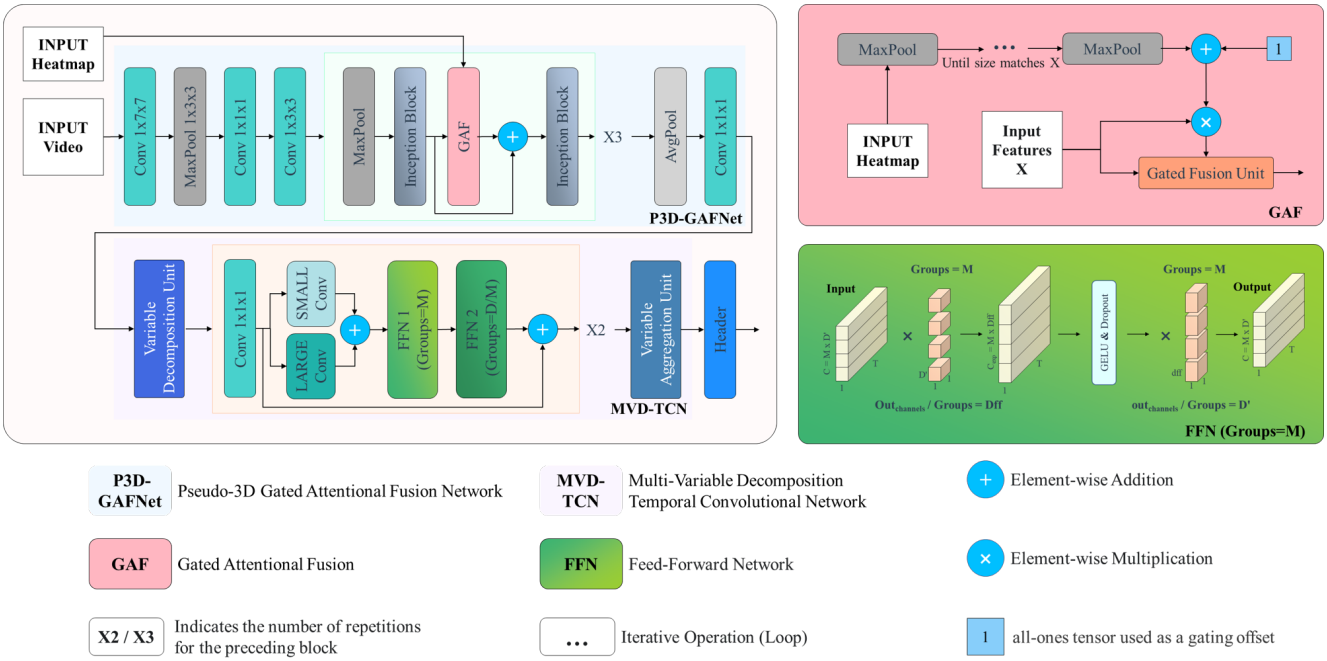

Figure 2. Overall architecture of DSTCNet. An input video is first processed by P3D-GAFNet to extract per-frame spatial features, which are then fed into MVD-TCN for temporal modeling and final gesture classification. Top-left heatmaps are repeatedly injected into each GAF module as spatial priors. This two-stage decoupled design enables efficient and robust spatio-temporal feature learning.

high-dimensional visual feature sequences in gesture recognition. Our work aims to fill this gap, offering a streamlined yet powerful alternative to Transformer hybrids.

2.3. Prior-guided Spatial Modeling

Efficient spatial representation is crucial for gesture recognition. 3D CNNs like I3D [3] model space-time jointly but are computationally expensive. In contrast, Pseudo-3D (P3D) convolutions [21] offer an efficient alternative by decoupling 3D kernels into spatial and temporal components, which also facilitates using 2D pretrained models. Our P3D-GAFNet follows this lightweight paradigm. However, an efficient backbone alone is not sufficient to handle the complexities of real-world gesture recognition. Beyond efficiency, guiding the model's attention is equally critical. Prior knowledge injection has proven to be a valuable strategy, with examples including using skeletal topology as priors in gesture recognition [13] or injecting semantic cues via knowledge distillation in video understanding [25]. We build on these insights by proposing a more direct mechanism that treats spatial priors as a first-class input, fusing them with visual features end-to-end. This design choice offers two key advantages: it avoids the computational overhead and complex pipeline of distillation-based methods, and its modality-agnostic nature allows it to be easily adapted to different types of prior information, yielding a simpler and more generalizable solution.

3. Method

3.1. Overview

We propose DSTCNet, a dynamic hand gesture recognition framework with a two-stage decoupled design (see Fig. 2). Unlike traditional approaches that couple spatial and temporal modeling in a single architecture, it separates the task into per-frame spatial feature extraction and subsequent temporal modeling of the feature sequence. This separation simplifies the network, enhances feature discriminability, and allows each stage to specialize in its respective task. This modular design philosophy is central to our framework, enabling targeted innovation within each component while maintaining overall architectural simplicity.

Given an input video clip $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times C}$, where T, H, W, and C denote temporal length, height, width, and number of channels, the data is first processed by P3D-GAFNet (Sec. 3.2). P3D-GAFNet operates on each frame, applying stacked pseudo-3D ($1\times k\times k$) convolutions to extract spatial features without temporal mixing, further augmented by GAF modules that incorporate heatmap priors to guide spatial attention. This yields a sequence of spatially refined features $\mathbf{F} \in \mathbb{R}^{T\times D}$, where D is the feature dimension. The sequence is then modeled by MVD-TCN (Sec. 3.3), which first partitions each feature vector into multiple distinct subspaces to reduce feature entanglement and enable parallel modeling of sub-dynamics. These

decomposed variables are then processed by modern convolutional blocks that combine parallel large- and small-kernel depthwise convolutions to capture multi-scale, long-range temporal dependencies. Finally, a decoupled feature interaction network models both intra-variable and inter-variable relationships, synthesizing a comprehensive representation of gesture dynamics in a purely convolutional and efficient framework. Temporal features are then aggregated via global average pooling and classified by a linear layer.

The entire framework is trained end-to-end. Its fully convolutional nature not only supports variable-length inputs and maintains temporal translation equivariance for improved robustness, but also culminates in a simple yet powerful design that achieves an optimal balance of accuracy and efficiency. For multi-modal inputs such as RGB and Depth, we adopt a late fusion strategy where a dedicated DSTCNet is trained for each modality, and their prediction scores are averaged during inference.

3.2. Spatial Feature Extraction with P3D-GAFNet

Effective spatial feature extraction from video frames is a cornerstone of gesture recognition. Mainstream approaches have historically faced a trade-off: 3D CNNs like I3D [3] jointly model spatio-temporal features but incur substantial computational costs, while 2D CNN backbones are efficient but temporally naive. To address this, we design P3D-GAFNet as our spatial backbone, drawing inspiration from the efficiency of Pseudo-3D architectures [21] and the multi-branch design of Inception networks [22].

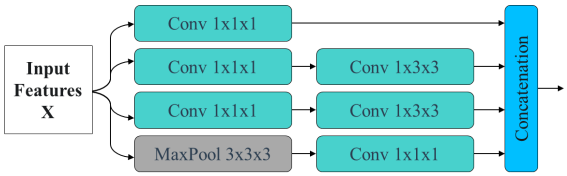

Figure 3. The structure of our Inception-style convolutional block.

The core of P3D-GAFNet is a stack of specialized convolutional blocks, as illustrated in Fig. 3. This Inception-style block processes the input through four parallel branches before their outputs are concatenated. To manage computational complexity, all branches typically first utilize $1 \times 1 \times 1$ convolutions for dimensionality adjustment. Following this, two branches apply an expressive $1 \times 3 \times 3$ convolution, a third branch proceeds with its $1 \times 1 \times 1$ output, and the fourth employs a max-pooling operation. Crucially, the use of two structurally similar convolutional branches encourages the learning of diverse and complementary representations, as each can specialize in distinct patterns due to random initialization.

However, even such sophisticated convolutional designs treat all intra-frame regions equally, making them vulnerable to distractions from complex backgrounds or irrelevant motion. To overcome this limitation, we introduce the GAF module, which injects spatial priors as guidance heatmaps. In our primary experiments, these heatmaps are generated from hand keypoint coordinates, where a 2D Gaussian kernel is placed at each keypoint location for every frame, as detailed in Section 4. Rather than applying this guidance only at the input layer, we strategically interleave GAF modules between stacks of convolutional blocks. This progressive refinement strategy ensures the spatial priors guide feature extraction at multiple semantic levels, preventing early guidance signals from being diluted in deeper layers.

Let $\mathbf{f}_t \in \mathbb{R}^{C' \times H' \times W'}$ be the visual feature map at time t, where C', H', and W' are the channel count, height, and width of the feature map at that specific layer, respectively, and $\mathbf{g}_t$ the corresponding guidance heatmap. To handle potentially different spatial resolutions, we first align $\mathbf{g}_t$ with $\mathbf{f}_t$ by applying max-pooling until their dimensions match. The GAF module then refines $\mathbf{f}_t$ in three stages. First, it enhances regions specified by the spatial prior to obtain an attention-modulated feature map:

$$\mathbf{f'}_t = \mathbf{f}_t \odot (\mathbf{g}_t + 1), \tag{1}$$

where $\odot$ denotes element-wise multiplication. Next, a learnable gating mechanism adaptively fuses the original and enhanced representations via a pointwise convolution:

$$\mathbf{f}_t^{\text{gate}} = \text{Conv}_{1\times1\times1}([\mathbf{f}_t, \mathbf{f'}_t]) \approx w_1 \cdot \mathbf{f}_t + w_2 \cdot \mathbf{f'}_t, \tag{2}$$

where $[\cdot, \cdot]$ denotes channel concatenation. Finally, a residual connection integrates the gated feature back into the main feature path to yield the refined output:

$$\mathbf{f}_t^{\text{out}} = \mathbf{f}_t + \mathbf{f}_t^{\text{gate}}. \tag{3}$$

Through this process, the GAF module adaptively balances original contextual features and prior-enhanced features, relying more on guidance when the hand is clearly visible and more on raw context when priors are unreliable.

By combining efficient pseudo-3D convolutions with this robust attention mechanism, P3D-GAFNet produces compact, noise-resilient spatial descriptors for each frame, forming the input sequence $\mathbf{F} \in \mathbb{R}^{T \times D}$ for the subsequent temporal network.

3.3. Temporal Modeling with MVD-TCN

After obtaining the spatial feature sequence $\mathbf{F} \in \mathbb{R}^{T \times D}$, the proposed MVD-TCN models its temporal dependencies in a purely convolutional yet highly expressive manner. Unlike Transformer-based methods [7, 11] that capture long-range context at the cost of quadratic complexity, or conventional TCNs [2] that remain efficient but suffer from limited effective receptive fields, MVD-TCN adopts a modern convolutional paradigm [6, 16, 18]. It first decomposes each high-dimensional feature vector into multiple subspaces, allowing the network to separately model latent sub-dynamics.

It then applies parallel large- and small-kernel convolutions enhanced by structural re-parameterization, enabling both long-range modeling and inference efficiency.

3.3.1 Multi-Variable Feature Decomposition

A high-dimensional feature vector (e.g., $D = 1024$) often entangles semantic components such as wrist posture and finger configuration. To better capture their fine-grained dynamics, we partition $\mathbf{F}$ along the feature dimension into M distinct variables (subspaces). This operation reshapes the feature sequence into a new tensor $\mathbf{F}'$:

$$\mathbf{F}' \in \mathbb{R}^{M \times (D/M) \times T} \tag{4}$$

This enables parallel temporal modeling in each distinct subspace, preserving their semantic purity. We opt for this simple, uniform partitioning as it is parameter-free and computationally trivial. Crucially, it creates parallel learning pathways without imposing complex, potentially biased structural priors that might be required by more sophisticated decomposition methods.

At the end of all temporal modeling stages, these subspaces, denoted as $\mathbf{F'}_i \in \mathbb{R}^{(D/M) \times T}$ for $i = 1, \ldots, M$, are adaptively fused. The aggregated feature is given by:

$$\mathbf{F}_{\text{agg}} = \sum_{i=1}^{M} \alpha_i \cdot \mathbf{F'}_i, \quad \alpha_i = \frac{\exp(w_i)}{\sum_{j=1}^{M} \exp(w_j)} \tag{5}$$

where $\mathbf{w} \in \mathbb{R}^M$ is a learnable parameter vector and α_i is the normalized weight for the i-th variable. This mechanism lets the model emphasize sub-dynamics most relevant to each gesture.

3.3.2 Modern Convolutional Block

The main body of MVD-TCN is composed of modern convolutional blocks, each designed to efficiently process the decomposed feature map $\mathbf{F}'$. As illustrated in the block detail of our overall framework (Fig. 2), each block executes a two-step process: first modeling temporal dependencies, then facilitating feature interaction across channels.

First, to capture temporal patterns at multiple scales, the block employs a dual-branch, depth-wise convolution. A large-kernel branch is responsible for modeling long-range dependencies by establishing a large ERF. Concurrently, a parallel small-kernel branch focuses on local, high-frequency details. This multi-branch structure enriches the gradient flow during backpropagation and is particularly beneficial for stabilizing large-kernel convolution training. Crucially, during inference, we leverage Structural Re-parameterization [5] to seamlessly merge both branches into a single, highly efficient convolution. This fusion is possible because its components are linear operations during inference, allowing their parameters to be folded into a single kernel and bias, which preserves full modeling capacity while significantly improving speed and memory locality.

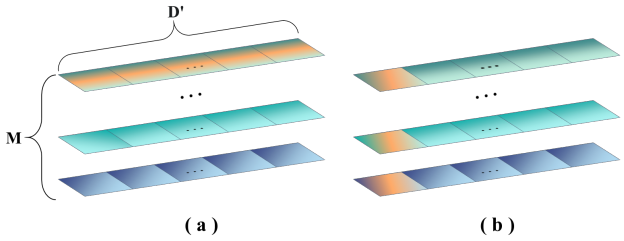

Figure 4. Illustration of our Decoupled Feature Interaction Network. (a) Intra-variable mixing processes each of the M variables in parallel to refine their individual features. (b) Inter-variable mixing exchanges information across the M variables to model their dependencies.

Following temporal modeling and layer normalization, our Feed-forward Network (FFN) enables decoupled feature interaction. As shown in Fig. 4, it performs two orthogonal mixing operations on decomposed variables. It begins with intra-variable mixing, where a grouped pointwise convolution (groups $= M$) processes the M variables in parallel (see Fig. 4(a)). This confines information flow to each variable's channels, enabling mining of individual sub-dynamics and forming specialized representations without interference. After feature rearrangement, it applies inter-variable mixing via another grouped pointwise convolution (groups $= D/M$), as in Fig. 4(b). This exchanges information across variables for each feature dimension, capturing dependencies and synergistic patterns among sub-dynamics. This structured two-stage design allows the block to model complex global dependencies—akin to attention mechanisms—while maintaining efficiency.

4. Experiments

4.1. Experimental Settings

Datasets. Our experiments are conducted on three public dynamic hand gesture recognition datasets, and in all cases, we strictly follow the official data splitting protocols for reproducibility and to ensure a fair comparison with prior work and thoroughly evaluate our model's performance and generalization capabilities.

NvGesture [20] is a large-scale benchmark for in-cabin human-computer interaction. It contains 25 gesture classes from 20 subjects, with 1,532 video samples captured by multiple sensors (RGB, Depth, IR). Its core challenges lie in handling multi-modal information and maintaining robustness against complex lighting and dynamic backgrounds.

Briareo [19] also targets in-cabin environments, comprising 12 gesture classes from 40 subjects. Its unique

bottom-up view introduces significant challenges from severe pose variations and sensor noise, making it a key benchmark for testing robustness to viewpoint changes.

EgoGesture [30] is a large-scale egocentric (first-person view) dataset with 83 gesture classes and over 24,000 samples. It is particularly challenging due to its high scene diversity, complex hand-object interactions, and motion blur caused by rapid movements.

Implementation Details and Metrics. All models are implemented in PyTorch and trained on an NVIDIA A40 GPU server. While our framework relies on an offline preprocessing step for guidance heatmap generation, the model itself is trained end-to-end. We use the AdamW optimizer with an initial learning rate of 1×10^{-4}, a batch size of 8, and a MultiStepLR scheduler for learning rate decay. Standard data augmentations, including random cropping, rotation, and scaling, are applied during training. The guidance heatmaps are generated from 2D hand keypoints extracted by a pre-trained detector from the MMPose toolbox [4]. For each frame, a heatmap is synthesized by placing a 2D Gaussian kernel (with σ=4) at each keypoint location and taking the element-wise maximum of all kernels. For multi-modal experiments, we employ a late fusion strategy by averaging prediction scores from individually trained models. For all competing methods, performance figures are directly quoted from their original publications. We evaluate performance using Top-1 Accuracy (%) and efficiency via Parameters (M) and MACs (G).

4.2. Main Performance Comparison

We conduct our primary evaluation on the NvGesture benchmark, a challenging and widely adopted dataset for robust gesture recognition. This enables a comprehensive and systematic comparison with representative methods spanning diverse technical paradigms.

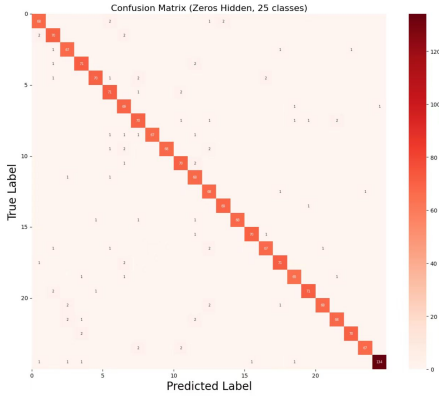

Figure 5. Confusion matrix of DSTCNet on NvGesture (Depth). The diagonal highlights high per-class accuracy.

Table 1. Comprehensive performance comparison on the NvGesture dataset.

Modality	Method	Accuracy (%) ↑
Depth	GestFormer [9]	80.21
	ConvMixFormer [11]	80.83
	Transformer [7]	83.00
	ResNeXt-101 [15]	83.82
	PreRNN [27]	84.40
	MTUT [1]	84.85
	MsMHA-VTN [8]	85.00
	MVTN [10]	85.21
	GPM [12]	85.50
	NAS1 [29]	86.10
	Ours	**88.33**
RGB	GestFormer [9]	75.41
	ConvMixFormer [11]	76.04
	Transformer [7]	76.50
	PreRNN [27]	76.50
	MVTN [10]	77.50
	ResNeXt-101 [15]	78.63
	MTUT [1]	81.33
	MsMHA-VTN [8]	81.42
	NAS1 [29]	**83.61**
	Ours	82.50
RGB + Depth	Transformer [7]	84.60
	MTUT [1]	85.48
	MMTM [14]	86.31
	NAS2 [29]	86.93
	NAS1+NAS2 [29]	88.38
	Ours	**89.86**

As summarized in Tab. 1, our model demonstrates consistently strong performance across both single-modality and multi-modality settings.

For the single-modality evaluation, our model achieves 88.33% accuracy on the Depth modality. This not only substantially surpasses the foundational Vision Transformer but also outperforms recent hybrid architectures such as MVTN and even the highly optimized NAS1. These results highlight the effectiveness of our purely convolutional framework, where the MVD strategy enables learning highly discriminative spatio-temporal representations from depth maps. On the RGB modality, our model attains 82.50% accuracy, further validating its robustness. As shown in the confusion matrix (Fig. 5), the model achieves high per-class accuracy, with minor confusion occurring primarily between semantically similar gestures.

For the multi-modal fusion evaluation, performance improves further. By integrating RGB and Depth information, our model achieves an accuracy of 89.96%, outperform-

ing the Transformer baseline (84.60%) and approaching the state-of-the-art accuracy reported by NAS-based methods. This demonstrates that our model can effectively exploit complementary multi-modal cues to achieve more robust gesture recognition.

4.3. Generalization Analysis

To further assess the generalization capability of our framework, we evaluate it on two additional public datasets with distinct characteristics: Briareo and EgoGesture.

Table 2. Performance comparison on the EgoGesture and Briareo datasets.

Dataset	Method	Accuracy (%) ↑	
		Color	Depth
EgoGesture	CatNet [26]	90.05	90.90
	MTUT [1]	92.48	91.96
	SeST [23]	93.20	93.35
	ResNeXt-101 [15]	**93.75**	94.03
	NAS1 [29]	93.31	94.13
	Ours	93.70	**94.26**
Briareo	Transformer [7]	90.60	92.40
	MsMHA-VTN [8]	98.15	94.44
	GestFormer [9]	94.44	96.18
	ConvMixFormer [11]	**98.26**	97.22
	MVTN [10]	97.69	97.92
	Ours	97.81	**98.10**

Briareo Dataset. As reported in Tab. 2, our model achieves a competitive accuracy of 98.10% on the Briareo dataset. This benchmark adopts a bottom-up viewpoint that introduces severe variations in hand shape, scale, and orientation, thereby challenging spatial feature extraction. Our strong performance, particularly against other methods under the Depth modality, indicates that P3D-GAFNet effectively mitigates viewpoint-induced distortions and enhances robustness to non-frontal perspectives.

EgoGesture Dataset. We then evaluate our model on the more challenging EgoGesture dataset. The results, summarized in Tab. 2, show that our framework maintains consistently competitive accuracy despite the dataset's complexity. EgoGesture is characterized by its egocentric viewpoint, cluttered backgrounds, and frequent hand-object interactions, which collectively introduce substantial visual ambiguities. The robustness of our framework under these conditions clearly demonstrates its adaptability to real-world scenarios and its ability to capture essential gesture dynamics despite strong environmental noise.

Overall Analysis. The consistently high performance across three representative datasets—covering fixed in-cabin views (NvGesture), bottom-up in-cabin views (Briareo), and egocentric interactions (EgoGesture)—demonstrates that our framework is not biased toward a single data distribution. Instead, it establishes a robust and generalizable paradigm for dynamic hand gesture recognition. Notably, the ability to maintain competitive results under diverse viewpoints, interaction styles, and environmental conditions highlights the effectiveness of our decoupled design in achieving reliable spatio-temporal modeling.

4.4. Model Efficiency Analysis

Table 3. Model efficiency comparison with state-of-the-art methods. MACs for our model are calculated on an input of 40 frames with a spatial resolution of 224×224. Results with * are reproduced by the authors.

Method	Parameters (M) ↓	MACs (G) ↓
NAS1 [29]	93.90	-
ResNeXt-101 [15]	52.28	74.44*
Transformer [7]	24.30	62.92
GestFormer [9]	24.08	60.40
MVTN [10]	19.55	60.22
ConvMixFormer [11]	13.57	59.98
Ours	**9.60**	**42.56**

DSTCNet emphasizes high computational efficiency while maintaining strong recognition performance. As shown in Tab. 3, our model requires only 9.60M parameters, corresponding to a 60.5% reduction compared to Transformer (24.30M) and a 50.9% reduction compared to MVTN (19.55M). This compact design substantially reduces storage and memory footprint, facilitating deployment in resource-constrained environments such as mobile devices or embedded systems.

The computational cost of DSTCNet is 42.56G MACs, which is significantly lower than other recent methods, including ConvMixFormer (59.98G MACs), GestFormer (60.40G MACs), and Transformer (62.92G MACs). This efficiency is achieved while maintaining a wide effective receptive field through large-kernel convolutions, allowing for effective aggregation of temporal and spatial features without increasing the computational burden.

Overall, DSTCNet establishes an efficient wide-ERF convolutional architecture that balances low computational cost and high representational capacity. This design provides a practical solution for real-time dynamic hand gesture recognition, combining both speed and accuracy for deployment in real-world scenarios.

Table 4. **Ablation study on the NvGesture Depth modality.** We evaluate the effectiveness of the spatial module (P3D-GAFNet), temporal module (MVD-TCN), and key internal designs including Gated Attentional Fusion (GAF), large-kernel convolutions, and Multi-Variable Decomposition (MVD). Results are reported as mean (standard deviation) over 3 independent runs.

#	Spatial Mod.	GAF	Temporal Mod.	Large Kernel	Multi-Var. Decomp. (M)	Accuracy (%) ↑
1	ResNet-18	N/A	Transformer	N/A	N/A	83.00 (± 0.14)
2	ResNet-18	N/A	MVD-TCN	✓	✓ (M=2)	86.52 (± 0.32)
3	P3D-GAFNet	✗	MVD-TCN	✓	✓ (M=2)	86.67 (± 0.21)
4	P3D-GAFNet	✓	MVD-TCN	✗	✓ (M=2)	84.38 (± 0.19)
5	P3D-GAFNet	✓	MVD-TCN	✓	✗ (M=1)	87.29 (± 0.12)
6	P3D-GAFNet	✓	MVD-TCN	✓	✓ (M=4)	87.19 (± 0.24)
7	P3D-GAFNet	✓	MVD-TCN	✓	✓ (M=8)	87.70 (± 0.18)
8	P3D-GAFNet	✓	MVD-TCN	✓	✓ (M=2)	88.33 (± 0.21)

4.5. Ablation Study

To dissect the contribution of each component in our framework, we conducted a series of controlled ablation experiments on the NvGesture dataset, as summarized in Tab. 4. For statistical robustness, each experiment was repeated 3 times with different random seeds. In each comparison, all other settings were kept identical to isolate the effect of the target modification, ensuring observed differences directly reflect the contribution of the modified component.

Temporal module. We begin with a standard ResNet-18 spatial backbone and compare a strong Transformer baseline (#1) against our proposed MVD-TCN (#2) as the temporal module. The results clearly indicate that MVD-TCN delivers significantly superior performance, validating the clear advantage of our modernized convolutional design for temporal modeling. This improvement can be mainly attributed to MVD-TCN's ability to capture long-range dependencies more effectively through large-kernel convolutions and multi-variable feature decomposition.

Spatial module. With MVD-TCN fixed, replacing the ResNet-18 backbone (#2) with our attention-equipped P3D-GAFNet (#8) yields a substantial accuracy gain. This confirms that P3D-GAFNet serves as a more powerful spatial feature extractor, likely due to its ability to emphasize relevant spatial regions and mitigate background noise through the guided attention mechanism. Such high-quality spatial features not only boost recognition accuracy but also enhance subsequent temporal modeling.

Internal components. Starting from the full model (#8), removing the GAF attention module (#3) leads to a noticeable accuracy decline, demonstrating the value of prior knowledge injection. Within MVD-TCN, discarding the large-kernel design (#4) or disabling our Multi-Variable Feature Decomposition strategy (#5) both cause significant degradation, underlining their importance as core temporal modeling elements. Notably, the performance drop from removing these temporal components is more severe in the presence of P3D-GAFNet, suggesting a synergistic relationship: high-quality spatial features are best exploited when both long-range modeling and disentangled sub-dynamics are preserved. This highlights the importance of co-designing spatial and temporal modules.

Effect of decomposition granularity. Varying the number of variables M (#5, #6, #7, #8) clearly and consistently reveals that overall performance peaks at $M = 2$, outperforming the non-decomposed case ($M = 1$, #5) and showing no additional gains at $M = 4$ or $M = 8$. We therefore adopt $M = 2$ as our default experimental setting, offering the absolute best possible trade-off between accuracy and computational efficiency. This strongly suggests that a moderate level of decomposition effectively captures the majority of meaningful temporal variations in dynamic hand gestures while avoiding fragmentation or insufficient modeling of distinct sub-dynamics.

5. Conclusion

We presented MVD-Net, a lightweight pure-convolutional framework (9.6M parameters) designed to efficiently capture long-range temporal dependencies. It achieves a superior accuracy-efficiency balance compared to larger models across multiple benchmarks. Beyond empirical success, this work demonstrates that modernized CNNs offer a preferable alternative to complex Transformers in resource-constrained scenarios. Future explorations will extend this paradigm to action detection and video segmentation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62272207 and 61772245, and by the Jiangxi Natural Science Foundation under Grant No. 20224ACB202009.

References

[1] Mahdi Abavisani, Hamid Reza Vaezi Joze, and Vishal M Patel. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1165–1174, 2019. 6, 7

[2] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. 2, 4

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3, 4

[4] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 6

[5] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13733–13742, 2021. 5

[6] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11963–11975, 2022. 2, 4

[7] Andrea D'Eusanio, Alessandro Simoni, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. A transformer-based network for dynamic hand gesture recognition. In *2020 International Conference on 3D Vision (3DV)*, pages 623–632. IEEE, 2020. 2, 4, 6, 7

[8] Mallika Garg, Debashis Ghosh, and Pyari Mohan Pradhan. Multiscaled multi-head attention-based video transformer network for hand gesture recognition. *IEEE Signal Processing Letters*, 30:80–84, 2023. 2, 6, 7

[9] Mallika Garg, Debashis Ghosh, and Pyari Mohan Pradhan. Gestformer: Multiscale wavelet pooling transformer network for dynamic hand gesture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2473–2483, 2024. 2, 6, 7

[10] Mallika Garg, Debashis Ghosh, and Pyari Mohan Pradhan. Mvtn: A multiscale video transformer network for hand gesture recognition. In *European Conference on Computer Vision*, pages 15–33. Springer, 2024. 1, 6, 7

[11] Mallika Garg, Debashis Ghosh, and Pyari Mohan Pradhan. Convmixformer-a resource-efficient convolution mixer for transformer-based dynamic hand gesture recognition. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6156–6166. IEEE, 2025. 1, 2, 4, 6, 7

[12] Vikram Gupta, Sai Kumar Dwivedi, Rishabh Dabral, and Arjun Jain. Progression modelling for online and early gesture detection. In *2019 International Conference on 3D Vision (3DV)*, pages 289–297. IEEE, 2019. 6

[13] Omar Ikne, Benjamin Allaert, and Hazem Wannous. Skeleton-based self-supervised feature extraction for improved dynamic hand gesture recognition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2024. 3

[14] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13289–13299, 2020. 6

[15] Okan Köpüklü, Ahmet Gunduz, Neslihan Kose, and Gerhard Rigoll. Real-time hand gesture detection and classification using convolutional neural networks. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–8. IEEE, 2019. 6, 7

[16] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2, 4

[17] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 1

[18] Donghao Luo and Xue Wang. Moderntcn: A modern pure convolution structure for general time series analysis. In *The twelfth international conference on learning representations*, pages 1–43, 2024. 2, 4

[19] Fabio Manganaro, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. Hand gestures for the human-car interaction: The briareo dataset. In *International Conference on Image Analysis and Processing*, pages 560–571. Springer, 2019. 5

[20] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4207–4215, 2016. 5

[21] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 3, 4

[22] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 4

[23] Xianlun Tang, Zhenfu Yan, Jiangping Peng, Bohui Hao, Huiming Wang, and Jie Li. Selective spatiotemporal features learning for dynamic gesture recognition. *Expert Systems with Applications*, 169:114499, 2021. 7

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[25] Guanhong Wang, Yang Zhou, Zhanhao He, Keyu Lu, Yang Feng, Zuozhu Liu, and Gaoang Wang. Knowledge-guided

pre-training and fine-tuning: Video representation learning for action recognition. *Neurocomputing*, 571:127136, 2024. 3

[26] Zhengwei Wang, Qi She, Tejo Chalasani, and Aljosa Smolic. Catnet: Class incremental 3d convnets for lifelong egocentric gesture recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 230–231, 2020. 7

[27] Xiaodong Yang, Pavlo Molchanov, and Jan Kautz. Making convolutional networks recurrent for visual sequence learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2018. 6

[28] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022. 2

[29] Zitong Yu, Benjia Zhou, Jun Wan, Pichao Wang, Haoyu Chen, Xin Liu, Stan Z Li, and Guoying Zhao. Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. *IEEE Transactions on Image Processing*, 30:5626–5640, 2021. 6, 7

[30] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia*, 20(5):1038–1050, 2018. 6