# Manta: Enhancing Mamba for Few-Shot Action Recognition of Long Sub-Sequence

**Wenbo Huang[1], Jinghui Zhang[1*], Guang Li[2], Lei Zhang[3], Shuoyuan Wang[4], Fang Dong[1], Jiahui Jin[1], Takahiro Ogawa[2], Miki Haseyama[2]**

[1]Southeast University, Nanjing 211189, Jiangsu, China
[2]Hokkaido University, Sapporo 060-0808, Hokkaido, Japan
[3]Nanjing Normal University, Nanjing 210023, Jiangsu, China
[4]Southern University of Science and Technology, Shenzhen 518055, Guangdong, China
wenbohuang1002@outlook.com, {jhzhang, jjin, fdong}@seu.edu.cn, {guang, ogawa, mhaseyama}@lmd.ist.hokudai.ac.jp, leizhang@njnu.edu.cn, claytonwang0205@gmail.com

## Abstract

In few-shot action recognition (FSAR), long sub-sequences of video naturally express entire actions more effectively. However, the high computational complexity of mainstream Transformer-based methods limits their application. Recent Mamba demonstrates efficiency in modeling long sequences, but directly applying Mamba to FSAR overlooks the importance of local feature modeling and alignment. Moreover, long sub-sequences within the same class accumulate intra-class variance, which adversely impacts FSAR performance. To solve these challenges, we propose a **M**atryoshka M**A**mba and Co**N**tras**T**ive Le**A**rning framework (**Manta**). Firstly, the Matryoshka Mamba introduces multiple Inner Modules to enhance local feature representation, rather than directly modeling global features. An Outer Module captures dependencies of timeline between these local features for implicit temporal alignment. Secondly, a hybrid contrastive learning paradigm, combining both supervised and unsupervised methods, is designed to mitigate the negative effects of intra-class variance accumulation. The Matryoshka Mamba and the hybrid contrastive learning paradigm operate in two parallel branches within Manta, enhancing Mamba for FSAR of long sub-sequence. Manta achieves new state-of-the-art performance on prominent benchmarks, including SSv2, Kinetics, UCF101, and HMDB51. Extensive empirical studies prove that Manta significantly improves FSAR of long sub-sequence from multiple perspectives.

## Introduction

Few-shot action recognition (FSAR) addresses the labeling reliance in data-driven training by classifying unseen actions from few video samples. This approach is widely used in real-world applications such as intelligent surveillance, video understanding, and health monitoring (Liu and Ma 2019; Reddy et al. 2022; Croitoru et al. 2021). Long video sub-sequences intuitively offer advantages in expressing the entire process of an action, like "Diving cliff", due to richer contextual information. In contrast, shorter sub-sequences may only cover partial actions, such as "Falling", "Running", and "Swimming". Despite this, research on the usage of long sub-sequences in FSAR remains unexplored.

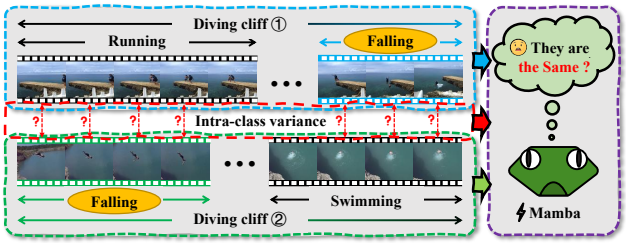



Figure 1: In two long sub-sequence examples of "Diving cliff", significant local features (highlighted as "Falling") occupy only small portions of the examples and are located at different points in the timeline. Additionally, the frame pairs from these examples exhibit large discrepancies in visual features. As the number of frames increases, intra-class variance gradually accumulates.

Mainstream Transformer-based methods (Vaswani et al. 2017) for FSAR are limited to processing short sub-sequences of around 8 frames due to their computational complexity. Recently, Mamba (Gu and Dao 2023; Dao and Gu 2024) has been applied to various tasks for its efficient long-sequence modeling capabilities without adding heavy computational overhead. Leveraging state space models (SSMs) (Gu et al. 2023), Mamba not only eliminates the complex attention mechanisms of Transformers but also flexibly manages the propagation and discarding of contextual information. However, the emphasis on global feature modeling by data-driven training in Mamba is misaligned with the extremely limited sample availability in FSAR.

Therefore, while applying Mamba to FSAR with long sub-sequences appears promising, it still faces two inherent challenges, as illustrated in Figure 1. ***Challenge 1: The absence of local feature modeling and alignment.*** Though inconspicuous, some local features are crucial for accurate recognition. In two examples of "Diving cliff", the core local features associated with "Falling" constitute only a small portion of the long sub-sequence, with the majority being secondary features. Focusing on global feature by Mamba often overlooks these critical local features, leading to po-

*Corresponding author.

tential misclassification in FSAR. Additionally, the core "Falling" features in different samples are not aligned temporally, and the absence of temporal alignment in Mamba exacerbates this issue, significantly degrading performance. ***Challenge 2: The intra-class variance accumulation of long sub-sequences.*** Influenced by factors such as shooting conditions or post-processing, frame pairs between different "Diving cliff" examples of long sub-sequences exhibit significant visual discrepancies, regardless of alignment. As the number of frames increases, intra-class variance gradually accumulates, making it more challenging to cluster samples of the same class and leading to possible misclassification.

Metric-based meta-learning is the mainstream paradigm in FSAR for efficacy and simplicity. After feature extraction, it embeds support samples into class prototypes for calculating distances between query samples, performing classification. Previous works directly apply explicit temporal alignment between sub-sequences (Cao et al. 2020; Xing et al. 2023a), inevitably ignoring local features. To improve this issue, recent works focus on the combination of global and local features, achieving satisfactory results (Perrett et al. 2021; Wang et al. 2022, 2023). However, we observe that they all utilize Transformer for short sub-sequence and are limited by complex calculation. In addition, the accumulation of intra-class variance is not severe due to the short sub-sequences, which are constantly overlooked. So far, solutions to the above challenges are absent.

Based on these observations, we propose the **M**atryoshka M**A**mba and Co**N**tras**T**ive Le**A**rning framework (**Manta**). Firstly, the Matryoshka Mamba employs multiple Inner Modules to enhance local features instead of directly modeling global feature. An Outer Module is designed for implicit temporal alignment by capturing dependencies of timeline between local features. Secondly, a hybrid contrastive learning paradigm, which simultaneously incorporates both supervised and unsupervised methods, is developed to mitigate the impact of intra-class variance accumulation. The Matryoshka Mamba and the hybrid contrastive learning paradigm operate in parallel branches within Manta to enhance Mamba for FSAR of long sub-sequence.

To the best of our knowledge, Manta is the first work to apply long sub-sequences and Mamba in FSAR. Our key contributions are threefold.

- We propose the Matryoshka Mamba for local feature modeling and alignment. The Inner Modules enhance local features from fragments of a long sub-sequence, while the Outer Module bidirectionally scans the entire sequence to perform implicit temporal alignment through fusion. The nesting of Inner Modules within the Outer Module makes the Matryoshka Mamba a more suitable model for FSAR of long sub-sequence.
- We design a hybrid contrastive learning paradigm for FSAR of long sub-sequence. Supervised contrastive learning is applied to labeled support samples, while an unsupervised method is used for unlabeled query samples. Subsequently, all samples are considered in an unsupervised manner. This approach enhances sample clustering and mitigates the negative impact of intra-class variance accumulation.
- Extensive experiments reveal that Manta achieves new state-of-the-art (SOTA) performance on several FSAR benchmarks, including SSv2, Kinetics, UCF101, and HMDB51. Further analysis highlights competitiveness of Manta, particularly for long sub-sequences.

## Related works

### Few-shot Action Recognition

The mainstream paradigm of FSAR is metric-based meta-learning with Transformer to temporal alignment. Among them, OTAM (Cao et al. 2020) employs dynamic time warping (DTW) algorithm to calculate sub-sequence similarities, aligning query and support samples. Then temporal relation is further emphasized, representative works are ITANet (Zhang et al. 2021), $T^2$AN (Li et al. 2022), and STRM (Thatipelli et al. 2022). To emphasize local features, fine-grained modeling is applied by TRX (Perrett et al. 2021), HyRSM (Wang et al. 2022), SloshNet (Xing et al. 2023a), and SA-CT (Zhang et al. 2023). Besides, additional information is introduced into the model, such as depth (Fu et al. 2020), optical flow (Wanyan et al. 2023), and motion information (Wang et al. 2023; Wu et al. 2022). Although remarkable performance was achieved, the above works are almost all based on short sub-sequences due to the computational complexity of Transformer architecture.

### Mamba Architecture

Recently, Mamba with SSM (Gu and Dao 2023; Dao and Gu 2024) has gained more attention because of its promising performance in modeling long sequences. More works are quickly moved to applying Mamba on vision tasks. Specifically, ViM (Zhu et al. 2024) shares a similar idea with ViT, integrating Mamba into the vision model. Inspired by ResNet (He et al. 2016), VMamba (Liu et al. 2024) constructs a hierarchical structure with Mamba. To improve efficiency, EfficientVMamba (Pei, Huang, and Xu 2024) is designed by imposing selective scan. However, they are all unable to align local features in FSAR.

### Contrastive Learning

In recent years, contrastive learning (He et al. 2020; Chen et al. 2020) receives increasing attention for its promising ability to learn generic representation from unlabeled samples. Subsequently, supervised contrastive learning (Khosla et al. 2020) makes full use of labels, accurately finding the positive and negative samples. Several works (Zheng et al. 2022; Gidaris et al. 2019; Su, Maji, and Hariharan 2020) point out that contrastive learning can serve as an auxiliary loss in few-shot learning, effectively alleviating the negative impact from intra-class variance. Hence, contrastive learning has the potential to solve the challenge of intra-class variance accumulation in FSAR.

## Methodology

### Problem Definition

According to previous works (Cao et al. 2020; Perrett et al. 2021), the dataset is divided into three non-overlapping parts

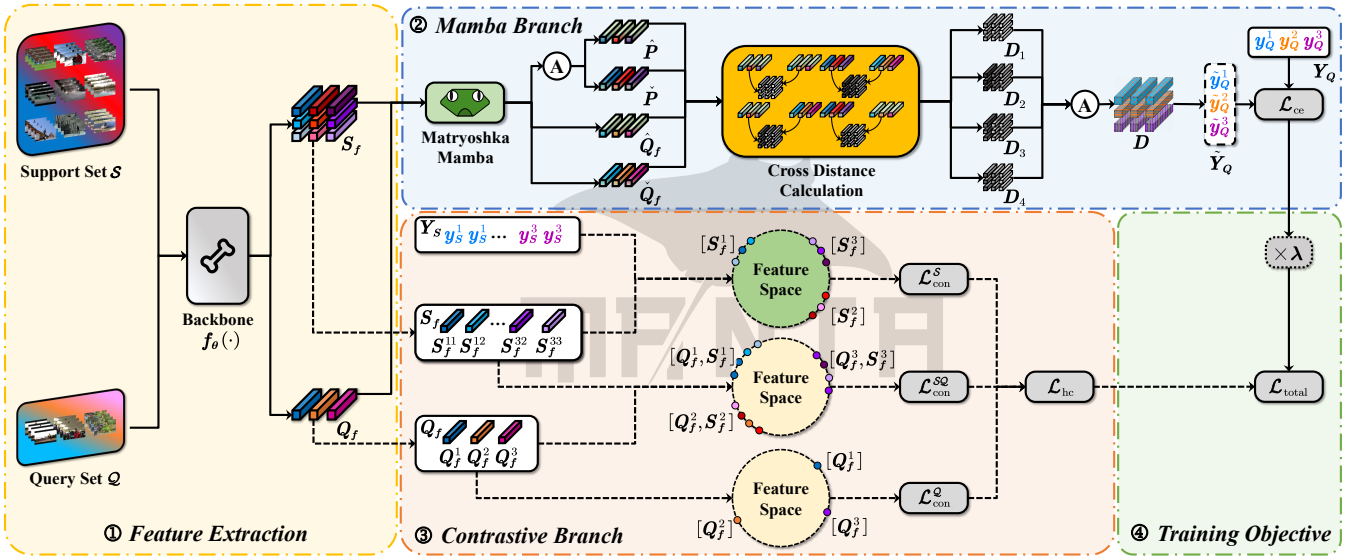



Figure 2: The overall architecture of the Matryoshka Mamba and Contrastive Learning framework (Manta) with four parts. To be specific, ① Feature Extraction with backbone extracts features from query and support. ② Mamba Branch with Matryoshka Mamba can emphasize local features and execute temporal alignment. ③ Contrastive Branch alleviates the accumulation of intra-class variance by hybrid contrastive learning. ④ Training Objective $\mathcal{L}_{\text{total}}$ is the loss combination of cross-entropy loss $\mathcal{L}_{\text{ce}}$ from ② Mamba Branch and contrastive loss $\mathcal{L}_{\text{hc}}$ from ③ Contrastive Branch. Notion Ⓐ means averaging calculation.

including training set $\mathcal{D}_{\text{train}}$, validation set $\mathcal{D}_{\text{val}}$, and test set $\mathcal{D}_{\text{test}}$ ($\mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{val}} \cap \mathcal{D}_{\text{test}} = \varnothing$). In each part, classifying unlabeled samples from query set $\mathcal{Q}$ into one class of support set $\mathcal{S}$ ($\mathcal{S} \cap \mathcal{Q} = \varnothing$) is the goal of FSAR. There is at least one labeled sample in each class of $\mathcal{S}$. In episodic training, a mass of few-shot tasks are randomly selected from $\mathcal{D}_{\text{train}}$. The $N$-way $K$-shot setting means that $\mathcal{S}$ in each task has $N$ classes and $K$ samples in each class.

## Overall Architecture

Figure 2 is an overview of Manta under 3-way 3-shot setting. A backbone is applied for feature extraction. In the Mamba branch, Matryoshka Mamba can enhance local feature modeling and alignment under various scales. Cross-entropy loss $\mathcal{L}_{\text{ce}}$ can be calculated by the distance between query and prototypes. In the contrastive branch, supervised and unsupervised paradigms work simultaneously, achieving hybrid contrastive learning loss $\mathcal{L}_{\text{hc}}$. The training objective $\mathcal{L}_{\text{total}}$ is the weighted combination of $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{hc}}$.

## Feature Extraction

Sub-sequences with $F$ frames are uniformly sampled from a video each time. The $k^{\text{th}}$ ($k = 1, \cdots, K$) support sample $S^{ck}$ in the $c^{\text{th}}$ ($c = 1, \cdots, N$) class of the support set $\mathcal{S}$ and the randomly selected query sample $Q^r$ ($r \in \mathbb{Z}^+$) from the query set $\mathcal{Q}$ are defined as follows:

$$\begin{aligned} S^{ck} &= \left[s_1^{ck}, \ldots, s_F^{ck}\right] \in \mathbb{R}^{F\times C\times H\times W}, \\ Q^r &= \left[q_1^r, \ldots, q_F^r\right] \in \mathbb{R}^{F\times C\times H\times W}. \end{aligned} \tag{1}$$

Notions applied are $F$ (frames), $C$ (channels), $H$ (height), and $W$ (width), respectively. $S^{ck}$ and $Q^r$ are sent into backbone $f_\theta(\cdot) : \mathbb{R}^{C\times H\times W} \mapsto \mathbb{R}^D$ for $D$-dimensional vectors

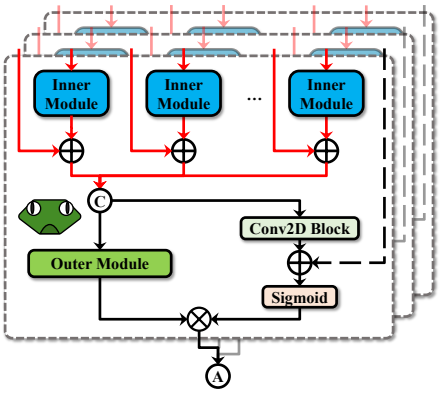



Figure 3: The structure of Matryoshka Mamba, ⊕, ⊗, and Ⓒ indicate element-wise addition, multiplication and concatenate operation. Conv2D Block has three 2D convolutions and a batch normalization layer. Red indicates local features while feature itself is dotted line.

$S_f^{ck}, Q_f^r \in \mathbb{R}^{F\times D}$:

$$\begin{aligned} S_f^{ck} &= \left[f_\theta\left(s_1^{ck}\right), \ldots, f_\theta\left(s_F^{ck}\right)\right], \\ Q_f^r &= \left[f_\theta\left(q_1^r\right), \ldots, f_\theta\left(q_F^r\right)\right]. \end{aligned} \tag{2}$$

## Mamba Branch

The structure of Matryoshka Mamba is shown in Figure 3. Multiple Inner Modules for local feature modeling are nested within an Outer Module for alignment, designed with Mamba-2 for high efficiency. Other models including

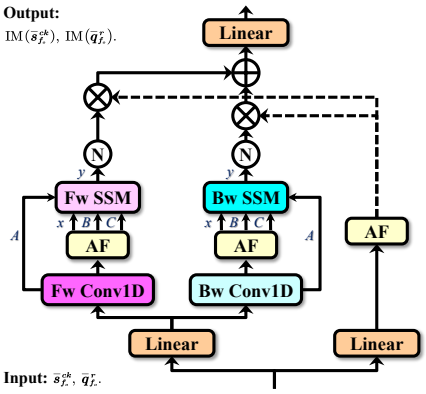



Figure 4: The structure of Inner Module based on Mamba-2, where Ⓝ, Fw, Bw, AF, and SSM refers to normalization, forward, backward, activation function, and state space model.

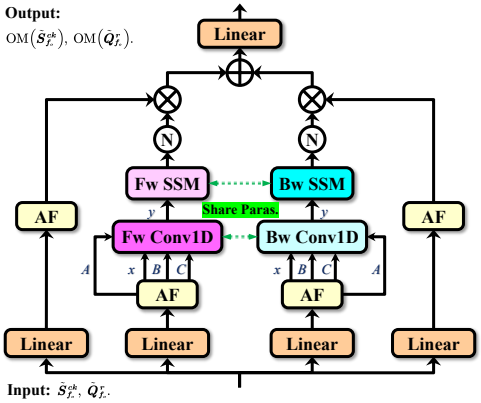



Figure 5: The bidirectional structure of Outer Module based on Mamba-2, decomposing the input at first. Two sub-branches share parameters.

Mamba-1 can also be utilized. The above structure is designed under a fixed scale. Building on this foundation, we extend the single scale to multiple scales for more comprehensive local feature modeling and alignment. The set $\mathcal{O}$ is defined as a hyper-parameter of multi-scale. The cardinality $|\mathcal{O}|$ denotes number of scales, while an element $o$ ($o \in \mathcal{O}$, $F \mid o$, $o < F$, $o = 2^{\alpha}$, and $\alpha \in \mathbb{Z}^{+}$) represents the frame count at this scale. For simplicity, we will use the subscript $o$ to indicate an arbitrary scale.

**Space State Models.** SSM in Mamba can effectively handle local feature modeling and alignment by flexibly propagating or discarding contextual information. It transforms input $x(t) \in \mathbb{R}^{L}$ to output $y(t) \in \mathbb{R}^{L}$ through hidden states $h(t) \in \mathbb{R}^{H}$. Linear ordinary differential equations are used for SSM description.

$$h'(t) = Ah(t) + Bx(t), y(t) = Ch(t). \tag{3}$$

$h'(t)$ is the derivative of $h(t)$. $A \in \mathbb{R}^{H \times H}$ is the state transition matrix while $B, C \in \mathbb{R}^{H}$ are projection parameters.

**Inner Module.** The feature of a long sub-sequence is divided into non-overlapping fragments, with Inner Modules enhancing local features from each fragment. As shown in Figure 4, the Inner Modules consist of two sub-branches, $\mathrm{IM}_{\mathrm{Fw}}(\cdot)$ and $\mathrm{IM}_{\mathrm{Bw}}(\cdot)$, which do not share parameters due to the differences in forward and backward local feature modeling. $\bar{s}^{ck}_{f_o}, \bar{q}^{r}_{f_o} \in \mathbb{R}^{o \times D}$ represent local feature tensors with length $L = D$ and hidden state $H = o$. The output is $\mathrm{IM}(\cdot) = \mathrm{Linear}\left[\mathrm{IM}_{\mathrm{Fw}}(\cdot) \oplus \mathrm{IM}_{\mathrm{Bw}}(\cdot)\right] \in \mathbb{R}^{F \times o}$. Concatenating ($\mathrm{C}[\ldots, \cdot, \ldots]$) each enhanced local feature and then adding with input, $\tilde{S}^{ck}_{fo}, \tilde{Q}^{r}_{fo} \in \mathbb{R}^{F \times D}$ can also be seen as tensors with length $L = F$ and hidden state $H = D$.

$$\begin{aligned} \tilde{S}^{ck}_{f_o} &= \mathrm{C}\left[\ldots, \mathrm{IM}\left(\bar{s}^{ck}_{f_o}\right) \oplus \bar{s}^{ck}_{f_o}, \ldots\right], \\ \tilde{Q}^{r}_{f_o} &= \mathrm{C}\left[\ldots, \mathrm{IM}\left(\bar{q}^{r}_{f_o}\right) \oplus \bar{q}^{r}_{f_o}, \ldots\right]. \end{aligned} \tag{4}$$

**Outer Module.** As illustrated in Figure 5, two sub-branches $\mathrm{OM}_{\mathrm{Fw}}(\cdot), \mathrm{OM}_{\mathrm{Bw}}(\cdot)$ with shared parameters are employed for bidirectional scanning since the forward and backward alignments are identical, capturing temporal dependencies for implicit alignment. After fusing them, the output is $\mathrm{OM}(\cdot) = \mathrm{Linear}\left[\mathrm{OM}_{\mathrm{Fw}}(\cdot) \oplus \mathrm{OM}_{\mathrm{Bw}}(\cdot)\right] \in \mathbb{R}^{F \times D}$. Inspired by C3-STISR (Zhao et al. 2022), we design learnable weights $w^{S}_{o}, w^{Q}_{o}$ for weighted averaging of various scales. Learnable weights are calculated from the input of Outer Module and feature itself. We define the Conv2D Block as $\mathrm{CB}(\cdot)$. The above calculation is written as

$$\begin{aligned} w^{S}_{o} &= \mathrm{Sigmoid}\left[\mathrm{CB}\left(\tilde{S}^{ck}_{f_o}\right) \oplus S^{ck}_{f}\right], \\ w^{Q}_{o} &= \mathrm{Sigmoid}\left[\mathrm{CB}\left(\tilde{Q}^{r}_{f_o}\right) \oplus Q^{r}_{f}\right]. \end{aligned} \tag{5}$$

Through combining with outputs of Outer Module, arbitrarily scaled $\mathring{S}^{ck}_{f_o}, \mathring{Q}^{r}_{f_o} \in \mathbb{R}^{F \times D}$ are obtained, referring to

$$\mathring{S}^{ck}_{f_o} = w^{S}_{o} \otimes \mathrm{OM}\left(\tilde{S}^{ck}_{f_o}\right), \ \mathring{Q}^{r}_{f_o} = w^{Q}_{o} \otimes \mathrm{OM}\left(\tilde{Q}^{r}_{f_o}\right). \tag{6}$$

The final outputs $\hat{S}^{ck}_{f}, \hat{Q}^{r}_{f} \in \mathbb{R}^{F \times D}$ of Matryoshka Mamba are averaged from all scales as

$$\hat{S}^{ck}_{f} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \mathring{S}^{ck}_{f_o}, \ \hat{Q}^{r}_{f} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \mathring{Q}^{r}_{f_o}. \tag{7}$$

**Prototype Construction.** The prototype of support $\hat{P}^{c}$ is constructed by the averaged paradigm as

$$\hat{P}^{c} = \frac{1}{K} \sum_{k=1}^{K} \hat{S}^{ck}_{f}. \tag{8}$$

**Cross Distance Calculation.** To further enhance temporal alignment, we apply cross-distance calculation. Specifically, $\check{P}^{c}, \check{Q}^{r}_{f}$ are the inversion of $\hat{P}^{c}, \hat{Q}^{r}_{f}$. Considering symmetrical alignment, a higher possibility of the same class is indicated by the smaller distances between tensors sharing the same superscript or the larger distances between tensors of various superscripts. Therefore, we take the reciprocal of those distances with various superscripts.

$$\begin{aligned} D_1 &= \left\|\hat{P}^{c} - \hat{Q}^{r}_{f}\right\|, \ D_2 = \left\|\check{P}^{c} - \check{Q}^{r}_{f}\right\|, \\ D_3 &= \left\|\hat{P}^{c} - \check{Q}^{r}_{f}\right\|^{-1}, \ D_4 = \left\|\check{P}^{c} - \hat{Q}^{r}_{f}\right\|^{-1}. \end{aligned} \tag{9}$$

The distance $D$ between query and the $c^{\text{th}}$ support is the average of the above four distances. Therefore, the model can predict label $\tilde{y}_Q^j \in \tilde{Y}_Q$ of query as

$$\tilde{y}_Q^j = \underset{c}{\operatorname{argmin}} (D), \; D = \frac{1}{4} \sum_{i=1}^{4} D_i. \tag{10}$$

Cross-entropy loss $\mathcal{L}_{\text{ce}}$ is calculated from the predicted label $\tilde{y}_Q^j$ and the ground truth $y_Q^j \in Y_Q$.

$$\mathcal{L}_{\text{ce}} = -\frac{1}{N} \sum_{j=1}^{N} y_Q^j \log \left( \tilde{y}_Q^j \right). \tag{11}$$

## Contrastive Branch

For alleviating the negative impact of intra-class variance accumulation, a hybrid contrastive learning paradigm with supervised and unsupervised methods is applied. Corresponding loss is formulated as

$$\mathcal{L}_{\text{con}} = -\log \frac{e^{\text{sim}(z, z^{\text{p}})/\tau}}{e^{\text{sim}(z, z^{\text{p}})/\tau} + \sum_{r=1}^{R} e^{\text{sim}(z, z_r^{\text{n}})/\tau}}, \tag{12}$$

Here, $\text{sim}(\cdot, \cdot)$ represents cosine similarity in the feature space, and $\tau$ denotes the temperature hyper-parameter. In the supervised paradigm applied to the support set, features with the same label are treated as positive samples $z^{\text{p}}$, while the other $R$ features with different labels are treated as negative samples $z^{\text{n}}$. In the unsupervised paradigm applied to the query set and all samples, there is no label guidance (the construction of $z^{\text{p}}$ and $z^{\text{n}}$ is detailed in the Supplementary Materials). The three contrastive losses including supervised $\mathcal{L}_{\text{con}}^{\mathcal{S}}$ and unsupervised $\mathcal{L}_{\text{con}}^{\mathcal{Q}}, \mathcal{L}_{\text{con}}^{\mathcal{SQ}}$ are combined to form the hybrid contrastive loss $\mathcal{L}_{\text{hc}}$.

$$\mathcal{L}_{\text{hc}} = \mathcal{L}_{\text{con}}^{\mathcal{S}} + \mathcal{L}_{\text{con}}^{\mathcal{Q}} + \mathcal{L}_{\text{con}}^{\mathcal{SQ}}. \tag{13}$$

## Training Objective

During the training stage, the above loss functions of two main branches supervise our Manta, the total loss $\mathcal{L}_{\text{total}}$ is

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{hc}}, \tag{14}$$

where $\lambda$ means the weight factor of $\mathcal{L}_{\text{ce}}$. In summary, Matryoshka Mamba enhances local feature modeling and alignment, while the hybrid contrastive learning paradigm alleviates the negative impact of intra-class variance accumulation. These combined operations make Manta a more suitable framework for FSAR of long sub-sequence.

# Experiments

## Experimental Configuration

**Data Processing.** Widely used benchmark datasets such as temporal-related SSv2 (Goyal et al. 2017), spatial-related Kinetics (Carreira and Zisserman 2017), UCF101 (Soomro, Zamir, and Shah 2012), and HMDB51 (Kuehne et al. 2011) are selected for proving the effectiveness of Manta. The sampling intervals are set to each 1 frame when decoding videos. According to the most common data split (Zhu and Yang 2018; Cao et al. 2020; Zhang et al. 2020), all datasets are divided into $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{test}}$ ($\mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{val}} \cap \mathcal{D}_{\text{test}} = \varnothing$). Tasks of FSAR aim to classify query samples $Q^r$ into corresponding classes of support set $\mathcal{S}$.

On the basis of TSN (Wang et al. 2016), frames are resized to $3 \times 256 \times 256$. $F$ frames per video are sequentially sampled each time. For simulating sub-sequences with various lengths, $F$ ($F \in [8, 128]$, $F \mid 2$) can be adjusted according to actual situation. In the regular setup, $F \geqslant 16$ can be seen as long sub-sequences. Data is augmented with $3 \times 224 \times 224$ random crops and horizontal flipping during training, while only the center crop is employed for testing. Due to SSv2 having many actions with horizontal direction such as "Pushing S from right to left"[1], horizontal flipping is absent in this dataset (Cao et al. 2020).

**Implementation Details and Evaluation Metrics.** We adopt two standard few-shot settings including 5-way 1-shot and 5-shot to conduct experiments. For a comprehensive comparison, ResNet-50 (He et al. 2016), ViT-B (Dosovitskiy et al. 2020), and VMamba-B (Liu et al. 2024) initialized with pre-trained weights on ImageNet (Deng et al. 2009) are served as the backbone. Features extracted are 2048-dimensional vectors ($D = 2048$).

Except for the larger SSv2 which requires 75,000 tasks training, other datasets utilize 10,000 tasks. An SGD optimizer with an initial learning rate of $10^{-3}$ is applied for training. The $\mathcal{D}_{\text{val}}$ determines hyper-parameters including multi-scale ($\mathcal{O} = \{1, 2, 4\}$), temperature ($\tau = 0.07$) and weight factor of loss ($\lambda = 4$). During the stage of testing, average accuracy across 10,000 random tasks of the testing set is reported. Most experiments are completed on a server with two 32GB NVIDIA Tesla V100 PCIe GPUs.

## Comparison with Various Methods

For a fair comparison with recent methods, we set the subsequence length as $F = 8$ and employ various backbones in this part. The average accuracy (↑ higher indicates better) is demonstrated in Table 1. Experiments on long sub-sequence are conducted in subsequent parts.

**ResNet-50 Methods.** Using the SSv2 dataset under 1-shot as an example, we observe that our Manta with ResNet-50 backbone improves the current SOTA method AMFAR from 61.7% to 63.4%. It is worth mentioning that AMFAR is a multimodal method with much heavier computational complexity than Manta. A similar improvement can also be observed in other datasets under various few-shot settings.

**ViT-B Methods.** In FSAR, ViT-B has fewer applications than ResNet-50. Methods using ViT-B tend to outperform because of its larger model capacity. For instance, in the 5-shot Kinetics dataset, the previous SOTA performance for RGB-based methods was achieved by MoLo. Following a similar trend with ResNet-50, Manta demonstrates superior performance, even surpassing the multimodal AMFAR.

**VMamba-B Methods.** As an emerging model, VMamba gains increasing attention for efficient feature extraction

[1]"S" denotes "something".

| Methods | Pre-Backbone | SSv2 | | Kinetics | | UCF101 | | HMDB51 | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1-shot | 5-shot | 1-shot | 5-shot | 1-shot | 5-shot | 1-shot | 5-shot |
| STRM (Thatipelli et al. 2022) | ImageNet-RN50 | N/A | 68.1 | N/A | 86.7 | N/A | 96.9 | N/A | 76.3 |
| SloshNet (Xing et al. 2023a) | ImageNet-RN50 | 46.5 | 68.3 | N/A | 87.0 | N/A | 97.1 | N/A | 77.5 |
| SA-CT (Zhang et al. 2023) | ImageNet-RN50 | 48.9 | 69.1 | 71.9 | 87.1 | 85.4 | 96.3 | 61.2 | 76.9 |
| GCSM (Yu et al. 2023) | ImageNet-RN50 | N/A | N/A | 74.2 | 88.2 | 86.5 | 97.1 | 61.3 | 79.3 |
| GgHM (Xing et al. 2023b) | ImageNet-RN50 | 54.5 | 69.2 | 74.9 | 87.4 | 85.2 | 96.3 | 61.2 | 76.9 |
| STRM (Thatipelli et al. 2022) | ImageNet-ViT | N/A | 70.2 | N/A | 91.2 | N/A | 98.1 | N/A | 81.3 |
| SA-CT (Zhang et al. 2023) | ImageNet-ViT | N/A | 66.3 | N/A | 91.2 | N/A | 98.0 | N/A | 81.6 |
| ⋆TRX (Perrett et al. 2021) | ImageNet-RN50 | 53.8 | 68.8 | 74.9 | 85.9 | 85.7 | 96.3 | 63.5 | 75.8 |
| ⋆HyRSM (Wang et al. 2022) | ImageNet-RN50 | 54.1 | 68.7 | 73.5 | 86.2 | 83.6 | 94.6 | 60.1 | 76.2 |
| ⋆MoLo (Wang et al. 2023) | ImageNet-RN50 | 56.6 | 70.7 | 74.2 | 85.7 | 86.2 | 95.4 | 67.1 | 77.3 |
| ⋆TRX (Perrett et al. 2021) | ImageNet-ViT | 57.2 | 71.4 | 76.3 | 87.5 | 88.9 | 97.2 | 66.9 | 78.8 |
| ⋆HyRSM (Wang et al. 2022) | ImageNet-ViT | 58.8 | 71.3 | 76.8 | 92.3 | 86.6 | 96.4 | 69.6 | 82.2 |
| ⋆MoLo (Wang et al. 2023) | ImageNet-ViT | 61.1 | 71.7 | 78.9 | 95.8 | 88.4 | 97.6 | 71.3 | 84.4 |
| ⋆TRX (Perrett et al. 2021) | ImageNet-VM | 56.9 | 71.5 | 76.2 | 87.2 | 88.1 | 97.0 | 66.7 | 78.5 |
| ⋆HyRSM (Wang et al. 2022) | ImageNet-VM | 58.6 | 71.7 | 76.6 | 92.4 | 86.8 | 96.5 | 70.2 | 82.6 |
| ⋆MoLo (Wang et al. 2023) | ImageNet-VM | 61.3 | 72.1 | 79.4 | 95.6 | 88.2 | 97.4 | 71.1 | 84.5 |
| AmeFu-Net (Fu et al. 2020) | ImageNet-RN50 | N/A | N/A | 74.1 | 86.8 | 85.1 | 95.5 | 60.2 | 75.5 |
| MTFAN (Wu et al. 2022) | ImageNet-RN50 | 45.7 | 60.4 | 74.6 | 87.4 | 84.8 | 95.1 | 59.0 | 74.6 |
| AMFAR (Wanyan et al. 2023) | ImageNet-RN50 | 61.7 | 79.5 | 80.1 | 92.6 | 91.2 | 99.0 | 73.9 | 87.8 |
| ⋆Lite-MKD (Liu et al. 2023) | ImageNet-RN50 | 55.7 | 69.9 | 75.0 | 87.5 | 85.3 | 96.8 | 66.9 | 74.7 |
| ⋆Lite-MKD (Liu et al. 2023) | ImageNet-ViT | 59.1 | 73.6 | 78.8 | 90.6 | 89.6 | 98.4 | 71.1 | 77.4 |
| ⋆Lite-MKD (Liu et al. 2023) | ImageNet-VM | 59.3 | 73.8 | 78.5 | 90.8 | 90.1 | 98.6 | 71.5 | 77.2 |
| Manta (Ours) | ImageNet-RN50 | **63.4** | **87.4** | **82.4** | **94.2** | **95.9** | **99.2** | **86.8** | **96.4** |
| Manta (Ours) | ImageNet-ViT | **66.2** | **89.3** | **84.2** | **96.3** | **97.2** | **99.5** | **88.9** | **96.8** |
| Manta (Ours) | ImageNet-VM | **66.1** | **89.1** | **84.4** | **96.2** | **96.9** | **99.4** | **89.1** | **96.6** |

Table 1: Comparison (↑ Acc. %) on ResNet-50 (ImageNet-RN50), ViT-B (ImageNet-ViT), and VMamba-B (ImageNet-VM). **Bold texts** denotes the global best results. From top to bottom, the whole table is divided into three parts including RGB-based, multimodal, and our Manta. In the first two parts, "⋆" represents our implementation with the same setting. "N/A" indicates not available in the corresponding publication. Underline texts serve as the local best results.

| IM | OM | $\mathcal{L}_{\text{hc}}$ | SSv2 | | Kinetics | |
|---|---|---|---|---|---|---|
| | | | 1-shot | 5-shot | 1-shot | 5-shot |
| ✗ | ✗ | ✗ | 46.3 | 64.5 | 70.9 | 86.5 |
| ✓ | ✗ | ✗ | 55.5 | 69.3 | 75.3 | 88.2 |
| ✗ | ✓ | ✗ | 55.3 | 69.4 | 75.0 | 88.0 |
| ✗ | ✗ | ✓ | 48.0 | 64.9 | 72.1 | 87.4 |
| ✓ | ✓ | ✗ | 63.8 | 88.0 | 82.8 | 94.6 |
| ✓ | ✗ | ✓ | 62.3 | 87.1 | 82.3 | 93.8 |
| ✗ | ✓ | ✓ | 61.9 | 86.7 | 81.9 | 93.4 |
| ✓ | ✓ | ✓ | **64.7** | **88.7** | **84.1** | **96.2** |

Table 2: Comparison (↑ Acc. %) of key components.

ability. Therefore, further comparison based on VMamba-B is also conducted in FSAR. Its performance is better than ResNet-50 because of a larger capacity. Our Manta with VMamba-B also achieves a remarkable improvement on performance, performing better than other VMamba-B based and even multimodal methods.

## Essential Components and Factors

All 1-shot experiments in this part are all trained and tested with ResNet-50 of long sub-sequences ($F = 16$).

**Key Components.** To verify the effect of key components, we split Manta into Inner Module (IM), Outer Module (OM), and hybrid contrastive learning loss ($\mathcal{L}_{\text{hc}}$) for testing. As indicated in Table 2, we have the following observation. Each key component improves the performance of the model. Therefore, applying the whole Manta brings the largest improvement by emphasizing local features, executing alignment, and reducing the negative impact of intra-class variance accumulation.

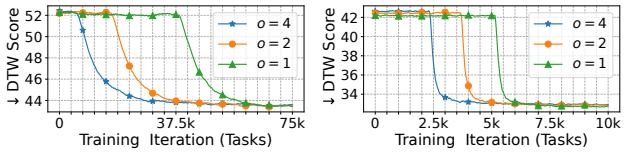
| Multi-Scale | SSv2 | | Kinetics | |
|---|---|---|---|---|
| | 1-shot | 5-shot | 1-shot | 5-shot |
| $\mathcal{O} = \{1\}$ | 63.3 | 87.2 | 81.6 | 92.9 |
| $\mathcal{O} = \{2\}$ | 63.4 | 87.0 | 81.3 | 93.2 |
| $\mathcal{O} = \{4\}$ | 63.2 | 87.3 | 81.4 | 93.1 |
| $\mathcal{O} = \{8\}$ | 63.0 | 86.9 | 81.1 | 92.7 |
| $\mathcal{O} = \{1, 2\}$ | 63.8 | 88.2 | 82.8 | 94.6 |
| $\mathcal{O} = \{1, 4\}$ | 64.1 | 87.9 | 83.4 | 94.5 |
| $\mathcal{O} = \{1, 8\}$ | 64.0 | 87.6 | 82.8 | 94.3 |
| $\mathcal{O} = \{2, 4\}$ | 63.9 | 88.4 | 83.2 | 94.3 |
| $\mathcal{O} = \{2, 8\}$ | 63.6 | 87.4 | 82.7 | 94.1 |
| $\mathcal{O} = \{4, 8\}$ | 63.6 | 87.3 | 82.8 | 94.0 |
| $\mathcal{O} = \{1, 2, 4\}$ | **64.7** | **88.7** | **84.1** | **96.2** |
| $\mathcal{O} = \{1, 2, 8\}$ | 64.5 | 88.5 | 83.8 | 96.0 |
| $\mathcal{O} = \{2, 4, 8\}$ | 64.4 | 88.2 | 83.6 | 95.8 |
| $\mathcal{O} = \{1, 2, 4, 8\}$ | 64.1 | 88.2 | 83.4 | 95.6 |

Table 3: Comparison (↑ Acc. %) of multi-scale design.

**Multi-Scale Design.** In Table 3, experiments with various hyper-parameter $\mathcal{O}$ are conducted for exploring how multi-scale design affect Manta. We find that emphasizing

local features can improve performance. When multi-scale is introduced, the improvement will be further expanded, as $|\mathcal{O}| = 3$ performs better than $|\mathcal{O}| = 2$ while $|\mathcal{O}| = 2$ defeats $|\mathcal{O}| = 1$. However, $|\mathcal{O}| = 4$ cannot further improve performance due to the redundant information. Therefor, we determine that $\mathcal{O} = \{1, 2, 4\}$ is the $\mathcal{O}$ setting without redundancy, performing better than other settings.

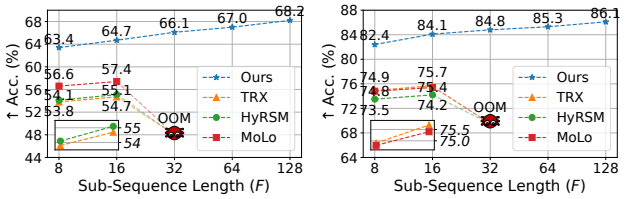
**Temporal Alignment.** These experiments are designed to reveal the trend of alignment from large to small scales. According to OTAM (Cao et al. 2020), we calculate DTW scores (↓ lower indicates better) from the shortest aligned paths of different scales. As illustrated in Figure 6, the DTW scores of two datasets all decline as training progresses, indicating that Manta learns temporal alignment gradually. In addition, the points of decline and convergence are sorted from early to late, proving that the alignment process of Manta is executed from coarse-grained to fine-grained.




(a) SSv2 (b) Kinetics

Figure 6: Alignment (↓ DTW scores) during training. $o = 4$, $o = 2$, and $o = 1$ mean large, middle, and small scales.

**Lengthening Sub-Sequence.** Through increasing $F$, we can get a long sub-sequence evaluation on three recent methods and our Manta. As shown in Figure 7, a remarkable performance improvement occurs when extending $F$ from 8 to 16. Any further extension of sub-sequence length will cause an "OOM" issue. These phenomena can be observed in all other methods while Manta performs stably even when the length of the sub-sequence is extremely lengthened to 128. We believe the promising performance is from long-sequence modeling and the special design of our Manta.




(a) SSv2 (b) Kinetics

Figure 7: Curves (↑ Acc. %, 1-shot setting) on long sub-sequence, where "OOM" means "out of memory".

**More Complex Tasks.** When processing long sub-sequences in FSAR, setting more classes ($N > 5$) in each task can enhance the complexity, evaluating the generalization of models. From the results demonstrated in Table 4, we can see that performance degrades due to more complex tasks. TRX, HyRSM, and MoLo suffer from a drastic decline in performance while our Manta can maintain a stable level. The above results prove that local feature emphasis, effective temporal alignment, and relieving intra-class variance accumulation are more needed when processing long sub-sequences in more complex tasks.

| Datasets | Methods | $N$-Way 1-Shot | | | | |
|---|---|---|---|---|---|---|
| | | 6 | 7 | 8 | 9 | 10 |
| SSv2 | TRX | 53.8 | 51.9 | 49.7 | 48.3 | 45.7 |
| | HyRSM | 53.6 | 50.4 | 48.1 | 45.2 | 42.9 |
| | MoLo | 55.2 | 53.0 | 50.8 | 48.5 | 45.1 |
| | Manta | **64.1** | **63.5** | **62.8** | **62.2** | **61.7** |
| Kinetics | TRX | 74.5 | 72.7 | 70.9 | 69.4 | 68.2 |
| | HyRSM | 73.1 | 70.2 | 67.5 | 64.8 | 62.5 |
| | MoLo | 74.3 | 71.5 | 72.7 | 67.1 | 64.2 |
| | Manta | **83.9** | **83.4** | **82.8** | **82.2** | **81.8** |

Table 4: Comparison (↑ Acc. %) under more complex tasks.

**Frame-Level Noise.** Frame-level noise means a sub-sequence mixed with multiple irrelevant frames ($F^*$). Conducting experiments with noise-level noise can simulate intra-class variance caused by authentic shooting conditions. Such a setting puts forward higher requirements for the robustness of models to face the uncertainty of samples. Results are revealed in Table 5. Intra-class variance accumulated by long sub-sequences has a negative impact. Several performance declines can be observed in TRX, HyRSM, and MoLo while our Manta is significantly less influenced. Therefore, we believe that our special design of Manta can improve the robustness and relieve the negative impact of intra-class variance accumulation.

| Datasets | Methods | Noisy Frame Numbers ($F^*$) | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 2 | 4 | 6 | 8 |
| SSv2 | TRX | 54.7 | 51.1 | 47.9 | 45.2 | 41.8 |
| | HyRSM | 55.1 | 51.8 | 48.9 | 45.7 | 42.1 |
| | MoLo | 57.4 | 54.2 | 51.4 | 49.1 | 45.8 |
| | Manta | **64.7** | **63.8** | **62.6** | **61.5** | **50.1** |
| Kinetics | TRX | 75.7 | 72.1 | 69.0 | 65.7 | 62.5 |
| | HyRSM | 74.2 | 70.9 | 67.6 | 64.7 | 61.5 |
| | MoLo | 75.4 | 72.5 | 69.4 | 66.1 | 62.9 |
| | Manta | **84.1** | **82.7** | **81.5** | **80.4** | **79.1** |

Table 5: Comparison (↑ Acc. %) with frame-level noise.

## Conclusion

In this paper, we propose Manta, specifically enhancing Mamba for FSAR of long sub-sequence. Manta emphasizes local feature modeling and alignment through Matryoshka Mamba and mitigates the negative impact of intra-class variance accumulation using a hybrid contrastive learning paradigm. Manta achieves new state-of-the-art performance across various benchmarks. Extensive studies demonstrate its competitiveness, generalization, and robustness, particularly for long sub-sequences. We hope our framework will inspire future research in FSAR.

## Acknowledgements

The authors would like to appreciate all participants of peer review and cloud servers provided by Paratera Ltd. Wenbo Huang sincerely thanks Bingxiao Shi (USTB), all family members, and TJUPT for the encouragement at an extremely difficult time. This work is supported by Frontier Technologies Research and Development Program of Jiangsu under Grant No. BF2024070; National Natural Science Foundation of China under Grants Nos. 62472094, 62072099, 62232004, 62373194, 62276063; Jiangsu Provincial Key Laboratory of Network and Information Security under Grant No. BM2003201; Key Laboratory of Computer Network and Information Integration (MOE, China) under Grant No. 93K-9; the Fundamental Research Funds for the Central Universities; the Excellent Ph.D Training Program (SEU); and JSPS KAKENHI Grant Nos. JP23K21676, JP24K02942, JP24K23849.

**Code** — https://github.com/wenbohuang1002/Manta

## References

Cao, K.; Ji, J.; Cao, Z.; Chang, C.-Y.; and Niebles, J. C. 2020. Few-shot video classification via temporal alignment. In *CVPR*.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*.

Croitoru, I.; Bogolin, S.-V.; Leordeanu, M.; Jin, H.; Zisserman, A.; Albanie, S.; and Liu, Y. 2021. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *ICCV*.

Dao, T.; and Gu, A. 2024. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *ICML*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Fu, Y.; Zhang, L.; Wang, J.; Fu, Y.; and Jiang, Y.-G. 2020. Depth guided adaptive meta-fusion network for few-shot video recognition. In *ACM MM*.

Gidaris, S.; Bursuc, A.; Komodakis, N.; Pérez, P.; and Cord, M. 2019. Boosting few-shot visual learning with self-supervision. In *ICCV*.

Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *ICCV*.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. In *COLM*.

Gu, A.; Johnson, I.; Timalsina, A.; Rudra, A.; and Ré, C. 2023. How to Train Your HiPPO: State Space Models with Generalized Orthogonal Basis Projections. In *ICLR*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. In *NeurIPS*.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: a large video database for human motion recognition. In *ICCV*.

Li, S.; Liu, H.; Qian, R.; Li, Y.; See, J.; Fei, M.; Yu, X.; and Lin, W. 2022. TA2N: Two-stage action alignment network for few-shot action recognition. In *AAAI*.

Liu, B.; Zheng, T.; Zheng, P.; Liu, D.; Qu, X.; Gao, J.; Dong, J.; and Wang, X. 2023. Lite-MKD: A Multi-modal Knowledge Distillation Framework for Lightweight Few-shot Action Recognition. In *ACM MM*.

Liu, K.; and Ma, H. 2019. Exploring background-bias for anomaly detection in surveillance videos. In *ACM MM*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*.

Pei, X.; Huang, T.; and Xu, C. 2024. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*.

Perrett, T.; Masullo, A.; Burghardt, T.; Mirmehdi, M.; and Damen, D. 2021. Temporal-relational crosstransformers for few-shot action recognition. In *CVPR*.

Reddy, R. G.; Rui, X.; Li, M.; Lin, X.; Wen, H.; Cho, J.; Huang, L.; Bansal, M.; Sil, A.; Chang, S.-F.; et al. 2022. MuMuQA: Multimedia Multi-Hop News Question Answering via Cross-Media Knowledge Extraction and Grounding. In *AAAI*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Su, J.-C.; Maji, S.; and Hariharan, B. 2020. When does self-supervision improve few-shot learning? In *ECCV*.

Thatipelli, A.; Narayan, S.; Khan, S.; Anwer, R. M.; Khan, F. S.; and Ghanem, B. 2022. Spatio-temporal relation modeling for few-shot action recognition. In *CVPR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*.

Wang, X.; Zhang, S.; Qing, Z.; Gao, C.; Zhang, Y.; Zhao, D.; and Sang, N. 2023. MoLo: Motion-augmented Long-short Contrastive Learning for Few-shot Action Recognition. In *CVPR*.

Wang, X.; Zhang, S.; Qing, Z.; Tang, M.; Zuo, Z.; Gao, C.; Jin, R.; and Sang, N. 2022. Hybrid relation guided set matching for few-shot action recognition. In *CVPR*.

Wanyan, Y.; Yang, X.; Chen, C.; and Xu, C. 2023. Active Exploration of Multimodal Complementarity for Few-Shot Action Recognition. In *CVPR*.

Wu, J.; Zhang, T.; Zhang, Z.; Wu, F.; and Zhang, Y. 2022. Motion-modulated temporal fragment alignment network for few-shot action recognition. In *CVPR*.

Xing, J.; Wang, M.; Liu, Y.; and Mu, B. 2023a. Revisiting the Spatial and Temporal Modeling for Few-shot Action Recognition. In *AAAI*.

Xing, J.; Wang, M.; Ruan, Y.; Chen, B.; Guo, Y.; Mu, B.; Dai, G.; Wang, J.; and Liu, Y. 2023b. Boosting Few-shot Action Recognition with Graph-guided Hybrid Matching. In *ICCV*.

Yu, T.; Chen, P.; Dang, Y.; Huan, R.; and Liang, R. 2023. Multi-Speed Global Contextual Subspace Matching for Few-Shot Action Recognition. In *ACM MM*.

Zhang, H.; Zhang, L.; Qi, X.; Li, H.; Torr, P. H.; and Koniusz, P. 2020. Few-shot action recognition with permutation-invariant attention. In *ECCV*.

Zhang, S.; et al. 2021. Learning implicit temporal alignment for few-shot video classification. In *IJCAI*.

Zhang, Y.; Fu, Y.; Ma, X.; Qi, L.; Chen, J.; Wu, Z.; and Jiang, Y.-G. 2023. On the Importance of Spatial Relations for Few-shot Action Recognition. In *ACM MM*.

Zhao, M.; Wang, M.; Bai, F.; Li, B.; Wang, J.; and Zhou, S. 2022. C3-stisr: Scene text image super-resolution with triple clues. In *IJCAI*.

Zheng, S.; et al. 2022. Few-shot action recognition with hierarchical matching and contrastive learning. In *ECCV*.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In *ICML*.

Zhu, L.; and Yang, Y. 2018. Compound memory networks for few-shot video classification. In *ECCV*.