Temporal Alignment-Free Video Matching for Few-shot Action Recognition

Supplementary Material

Datasets. We conduct experiments on four benchmark datasets for Few-Shot Action Recognition (FSAR): HMDB51 [15], Kinetics [5], UCF101 [26], and SSv2-Small [10]. For HMDB51, we adopt the settings from ARN [33], using 31, 10, and 10 classes for training, validation, and testing, respectively. Kinetics is adapted to the few-shot setting following previous works [3, 36], with 64, 12, and 24 categories assigned to training, validation, and testing. For UCF101, we follow the ARN [33] settings, using 70, 10, and 21 classes for training, validation, and testing. Lastly, for SSv2-Small, we use the split settings from CMN [35], with 64, 12, and 24 categories for training, validation, and testing.

Implementation Details. For a fair comparison with existing works, we follow existing protocols [3, 16, 29, 35]. We use ResNet-50 [11] and ViT-B [8] as a backbone network, both pre-trained on ImageNet [7]. The backbone processes each video by taking 8 uniformly sampled frames at a resolution of 224×224 as input (*i.e.*, $T = 8$). As in prior studies, data augmentation techniques such as random cropping and color jittering are applied [28, 29, 32]. Under the ResNet, the number of learnable pattern is The number of learnable pattern tokens is listed in Tab. 6. For the optimization, we adopt SGD to train our model for 10,000 iterations on all datasets. In the many-shot scenarios, we utilize the prototype concept [25], following MoLo [29]. We measure the performance of our model as the average results over 10,000 randomly sampled episodes. All experiments are conducted on RTX A6000 GPUs with Pytorch Cuda amp.

Table 6. The number of pattern tokens.

	ResNet		ViT	
	1-shot	5-shot	1-shot	5-shot
HMDB51	60	70	50	60
Kinetics	60	80	80	80
UCF101	60	80	70	70
SSv2-Small	-	-	50	80