Embedding Adaptation Network with Transformer for Few-Shot Action Recognition

Rongrong Jin RRONGJIN@STU.XMU.EDU.CN
Xiao Wang XIAOWANG@STU.XMU.EDU.CN
Guangge Wang GUANGGEW@STU.XMU.EDU.CN
Yang Lu LUYANG@XMU.EDU.CN
Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen, China

Hai-Miao Hu FRANK0139@163.COM
Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing, China

Hanzi Wang* HANZI.WANG@XMU.EDU.CN
Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen, China
* *Corresponding Author*

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

Few-shot action recognition aims to classify novel action categories using a few training samples. Most current few-shot action recognition methods via episodic training strategy mainly use the same normalization method to normalize feature embeddings, leading to limited performance when the batch size is small. And some methods learn feature embeddings individually without considering the whole task, neglecting important interactive information between videos in the current episode. To address these problems, we propose a novel embedding adaptation network with Transformer (EANT) for few-shot action recognition. Specifically, we first propose an improved self-guided instance normalization (SGIN) module to adaptively learn class-specific feature embeddings in an input-dependent manner. Built upon the learned feature embeddings, we design a Transformer-based embedding learning (TEL) module to learn task-specific feature embeddings by fully capturing rich information cross videos in each episodic task. Furthermore, we utilize semantic knowledge among all sampled training classes as additional supervisory information to improve the generalization ability of the network. By this means, the proposed EANT can be highly effective and informative for few-shot action recognition. Extensive experiments conducted on several challenging few-shot action recognition benchmarks show that the proposed EANT outperforms several state-of-the-art methods by a large margin.

Keywords: Few-shot Learning, Action Recognition, Embedding Adaptation, Transformer, Semantic Knowledge.

1. Introduction

Action recognition is an essential task in the field of computer vision with many practical applications. Deep learning-based action recognition methods (Wang et al. (2016); Xie et al.

(2018); Tran et al. (2018); Zhou et al. (2021); Wu et al. (2021); Dave et al. (2022)) have made significant progress due to the remarkable ability of feature representation and the availability of large amounts of annotated training data. However, the success is established on the intensive labor to annotate a large number of video data manually. In contrast, humans can recognize a new class of object with only one or a few samples. Thus, a promising direction has emerged in tackling the challenging few-shot learning (FSL) problem (Vinyals et al. (2016); Snell et al. (2017); Xiao et al. (2020)). FSL aims to train a classifier with the ability to classify unseen classes from a few annotated samples. To deal with this problem, many meta-learning based methods (Vinyals et al. (2016); Finn et al. (2017)) propose to learn a transferable model that can quickly adapt to the new target tasks. In recent years, metric-learning based methods (Sung et al. (2018); Ye et al. (2020); Xiao et al. (2020)) have been widely used for few-shot image classification. They propose to learn a similarity metric to calculate the distance between the query and support samples. Although these methods have achieved great success in few-shot image classification, limited work has been done on the task of few-shot video classification.

Generally, it still presents enormous challenges to classify novel video action categories given only a few labeled video samples. Compared to two-dimensional images, videos usually contain hundreds of frames, which involve complex higher-dimensional temporal information and different motion scenes. Besides, actions are usually performed at different speeds and occur at various times. In addition, deep learning-based methods with scarce labeled training data may suffer from limited scalability and are prone to overfitting, failing to learn good feature representations. To address these issues, current few-shot action recognition methods (Bishay et al. (2019); Cao et al. (2020); Zhang et al. (2020); Perrett et al. (2021); Zhang et al. (2021)) mainly adopt the metric-based meta-learning paradigm (Vinyals et al. (2016)) for its simplicity and effectiveness. It first randomly samples support sets and query sets episodically and then trains a classifier to compute the classification loss by calculating the distances between the query and support samples in each training episode.

Despite impressive progress have been achieved, these few-shot action recognition methods still encounter two problems. First, some of them (Bishay et al. (2019); Perrett et al. (2021)) propose to introduce effective temporal aggregation methods to combine feature vectors into video-level feature embeddings in an episodic task. These methods ignore the ordering of frames and may not capture temporal information between adjacent frames, failing to learn effective feature representations. Besides, they always use the same normalization method to normalize feature embeddings, where accurate estimation of normalization parameters may not be possible when the batch size is small, degrading the performance of the models to some extent. Second, some of them (Zhang et al. (2020); Cao et al. (2020); Zhang et al. (2021)) focus on designing temporal alignment modules to match variable length videos. Nevertheless, they encode features for each video individually and preserve substantial restrictions on the temporal ordering, ignoring discriminative interactive information between videos in each episode. As a result, the learned feature representations preserve equal impact on different episodic tasks, which are task-agnostic and may overlook the discrimination needed in the current task. Moreover, they only focus on visual features while neglecting the semantic knowledge in the labeled training data. We claim that the semantic knowledge can be utilized as auxiliary semantic information to improve the discriminability and informativeness of the learned feature representations.

To address these problems, we thus propose a novel embedding adaptation network with Transformer (EANT) for few-shot action recognition. Specifically, to address the first problem, we propose to improve discriminative feature representations learning for videos from a new perspective. We propose a new self-guided instance normalization (SGIN) module to enhance discriminative feature embeddings learning instead of simply aggregating the feature vectors into a single feature embedding. The SGIN module preserves the temporal information between frames and enables the model to focus on specific objects and their corresponding actions in a video. Built upon the learned class-specific feature embeddings, we design a Transformer-based embedding learning (TEL) module to solve the second problem, which jointly considers all the support and query features in an episodic task to customize the feature embeddings towards the specific task. In addition, we make full use of the class labels of the training data to discriminate the class-similarity among the support-query pairs, which is helpful to separate samples from different categories, especially for similar categories. By this means, the proposed EANT can achieve class-specific and task-specific discriminability enhancement within the entire task, yielding discriminative feature embeddings with better generalization ability.

In summary, the main contributions of this paper are three-fold.

- We propose a novel self-guided instance normalization module to adaptively enhance discriminative feature embeddings learning for videos in an input-dependent manner, obtaining class-specific and discriminative feature embeddings.
- We design a Transformer-based embedding learning module to capture interactive information cross videos in each episodic task, yielding task-specific feature embeddings for different tasks.
- We further leverage the semantic knowledge of the training class labels to distinguish fine-grained action categories. We conduct extensive experiments on three widely used challenging benchmarks to demonstrate that the proposed EANT achieves state-of-the-art experimental results.

2. Related Work

Few-shot action recognition. Few-shot action recognition has drawn growing attention in recent years. Current methods that achieve state-of-the-art performance mainly focus on metric-based learning. CMN (Zhu and Yang (2018)) designs a memory network structure that uses a key-value memory network paradigm to encode feature representations. To the best of our knowledge, it is the first work proposed for few-shot action recognition. Embodied Learning (Fu et al. (2019)) introduces a virtual dataset as a benchmark and proposes a video segment augmentation method to synthesize new videos. ProtoGAN (Dwivedi et al. (2019)) utilizes a conditional GAN to generate additional samples for novel categories. AMeFu-Net (Fu et al. (2020)) proposes to fuse depth information with visual information by the temporal asynchronization augmentation mechanism to synthesize new instance features. They all use auxiliary methods to generate additional samples. Instead, ARN (Zhang et al. (2020)) focuses on capturing both short-range and long-range temporal relations with the permutation-invariant attention. ITANet (Zhang et al. (2021)) proposes

an implicit temporal alignment strategy for video matching. TRX (Perrett et al. (2021)) leverages the CrossTransformer attention mechanism to construct query-specific class prototypes. HyRSM (Wang et al. (2022)) proposes a hybrid relation module and a set matching metric to improve the transferability of embedding. Unlike these previous methods, our EANT improves the discriminability of feature embeddings by learning class-specific and task-specific representations with better generalization ability.

Instance normalization. Batch normalization (BN) (Ioffe and Szegedy (2015)) has been successfully used to improve the optimization performance and accelerate the training process. Since then, a lot of variant normalization methods have been proposed, such as instance normalization (Ulyanov et al. (2016)) and adaptive normalization (Xu et al. (2019)). Instance normalization (IN) has been applied in style transfer with surprising effectiveness. Adaptive instance normalization (AdaIN) is a simple extension to IN, which can adjust the normalization parameters of the content input features to align with those of the style input features. These normalization methods mentioned above are primarily used for the image classification task. Inspired by this, we propose to learn normalization parameters from the original input video features adaptively, and then use these parameters to promote the training process of few-shot action recognition.

Transformer. Transformer is first proposed by Vaswani et al. (2017), which leverages the self-attention mechanism for natural language processing (NLP) (Bahdanau et al. (2015)) and shows impressive power in language translation. Compared to convolutions, Transformer can capture long-range context information and dynamically adjust weights according to the input. Since then, many works have adapted the merits of Transformer in computer vision tasks. For instance, FEAT (Ye et al. (2020)) employs a Transformer to obtain discriminative image feature embeddings via the self-attention mechanism. ViT (Dosovitskiy et al. (2021)) applies a pure Transformer structure to the image classification task and achieves state-of-the-art performance. TimeSformer (Bertasius et al. (2021)) adapts the standard Transformer architecture to learn spatial-temporal features from a sequence of frames. In this paper, we design a Transformer-based embedding learning module for few-shot action recognition, taking the vital contextual cues from different videos into consideration and obtaining refined feature embeddings that are permutation invariant and task-specific.

3. Proposed Method

3.1. Problem Formulation

Following the setting of few-shot action recognition (Cao et al. (2020)), there is a training set $\mathcal{D}_{train} = \{(v_i, c_i)|c_i \in \mathcal{C}_{train}\}$ and a testing set $\mathcal{D}_{test} = \{(v_i, c_i)|c_i \in \mathcal{C}_{test}\}$, where v_i and c_i denote the i-th video and its corresponding class label, respectively. The categories of the $\mathcal{D}_{train}$ and the $\mathcal{D}_{test}$ are disjoint, i.e., $\mathcal{C}_{train} \cap \mathcal{C}_{test} = \varnothing$. The goal of few-shot action recognition is to train a model on the $\mathcal{D}_{train}$. Then, the trained model is utilized to recognize novel categories in $\mathcal{D}_{test}$ given only few labeled samples per action class. Concretely, we randomly sample a great number of M-way K-shot tasks from the $\mathcal{D}_{train}$, and then we use them to train the model adopting the episodic training strategy (Sung et al. (2017); Zhu and Yang (2018)). In an M-way K-shot task, we sample M classes each with K videos

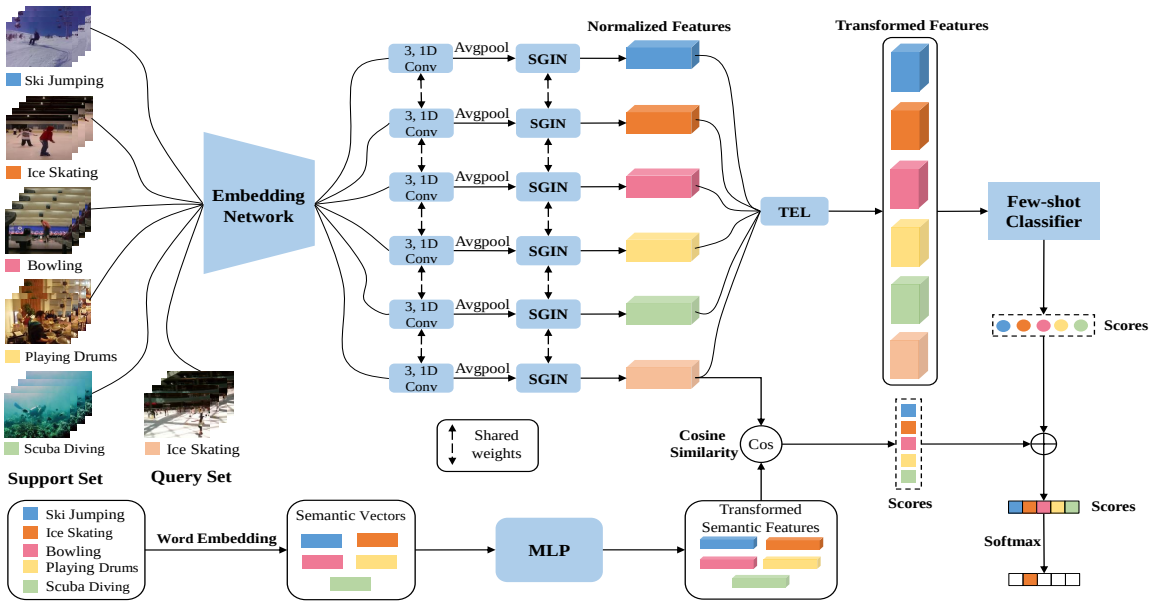


Figure 1: The overview of EANT. Given an episode of videos, an embedding network is first employed to extract their feature maps. A channel-wise 1D convolution is applied and then followed by the self-guided instance normalization module to generate class-specific features. After that, the Transformer-based embedding learning module is used to customize the normalized features towards the target task. Finally, a few-shot classifier is utilized to compute the classification scores, combining with the visual-semantic similarity scores to make the final prediction.

from $\mathcal{D}_{train}$ at random to form a support set $\mathcal{S}$, and the rest videos of the M classes are sampled to form a query set $\mathcal{Q}$. Following Zhu and Yang (2018), the query set composes of only one sample in each episode.

3.2. Embedding Adaptation Network with Transformer

Framework overview. The overview of the proposed EANT is illustrated in Figure 1. The input consists of the query and support videos. We first employ an embedding network (e.g., ResNet-50) to extract the feature embeddings for each input video frames. Then, we propose a self-guided instance normalization (SGIN) module to perform discriminative feature representations learning enhancement, obtaining class-specific and discriminative normalized feature embeddings. After that, we design a Transformer-based embedding learning (TEL) module to customize the feature embeddings towards a specific task. In this way, we can acquire transformed feature embeddings that are task-specific. All the transformed support features and query feature are fed into a few-shot classifier, and we get similarity scores between the query feature and support features. Besides, we use Word2Vec (Mikolov et al. (2013)) to get semantic vectors for the training class labels and then we employ a two-layer multi-layer perception (MLP) to map the semantic vectors to

the visual space. We compute the cosine similarity scores between the query feature and the transformed semantic features. Finally, we combine the two scores and apply a softmax layer on the combined distance similarities to obtain the predicted probabilities. The action category with the highest probability is selected as the final prediction.

Self-guided instance normalization module. To enhance effective feature embeddings learning for few-shot action recognition, we introduce an improved normalization design from a new perspective. Current video classification works always use batch normalization (BN) (Ioffe and Szegedy (2015)) to normalize feature embeddings. Despite the significance of BN has been demonstrated in many previous works, it tends to bring noise during the computation of normalization parameters when the batch size is small, thus leading to limited performance. Many normalization methods (Ulyanov et al. (2016); Xing et al. (2019)) have been proposed to alleviate this issue. However, the performance improvement of these methods is not obvious when applying them to videos since they are primarily designed for images. Directly applying them to the few-shot action recognition task may achieve limited performance, given the fact that videos have complex spatial-temporal structures and contain crucial sequence information.

Unlike the previous methods, we propose a self-guided instance normalization (SGIN) module to generate class-specific and discriminative features. Specifically, given the input features $f \in \mathbb{R}^{N\times T\times C\times H\times W}$ derived from the embedding network, where N is the batch size. T and C denote the temporal dimension and the feature channels, respectively. H and W represent the spatial size. We first reshape f from $[N, T, C, H, W]$ to $[NT, C, H, W]$ and apply the channel-wise 1D convolution layer on the T dimension to learn the temporal information. Then, we use the average pooling to aggregate the spatial information, and we can obtain the feature maps $x \in \mathbb{R}^{NT\times C\times 1\times 1}$. The SGIN module receives the feature maps x as input, and adaptively performs normalization on the original input feature maps:

$$f(x) = g_\gamma(x) \cdot \frac{x - \mu(x)}{\sigma(x)} + g_\beta(x), \tag{1}$$

where $g_\gamma(\cdot)$ and $g_\beta(\cdot)$ are both three learnable convolution layers with the kernal size of 1×1, respectively. Each convolution layer is followed by a batch normalization (BN) layer and a LeakyReLU activation function except the last convolution layer is followed by a sigmoid function. The $g_\gamma(\cdot)$ and $g_\beta(\cdot)$ are jointly trained with the whole network and the outputs are utilized as the scale and shift parameters to perform normalization. The $\mu_{n,t}(x)$ and $\sigma_{n,t}(x)$ are computed along the C dimension:

$$\mu_{n,t}(x) = \frac{1}{C}\sum_{c=1}^{C} x_{n,t,c}, \tag{2}$$

$$\sigma_{n,t}(x) = \sqrt{\frac{1}{C}\sum_{c=1}^{C}(x_{n,t,c} - \mu_{n,t}(x))^2 + \epsilon}, \tag{3}$$

where ϵ denotes a small positive constant. The SGIN module adaptively learns the normalization parameters from the input feature maps by using three learnable convolution layers. It then leverages the learned parameters to conduct normalization on the input itself, generating features X that are class-specific and discriminative.

Transformer-based embedding learning module. Some current metric-based meta-learning methods (Zhang et al. (2020); Cao et al. (2020); Zhang et al. (2021)) mainly focus on designing temporal alignment module to match segment embeddings. These methods take the long-term temporal information into consideration and retain temporal information between frames. However, they preserve strong restrictions on the temporal ordering and the discriminative interaction cross videos in each episode are ignored, since each video is processed independently during the training process.

Different from the above methods, we design a Transformer-based embedding learning (TEL) module to jointly extract the long-range contextual information needed by the specific action recognition task. Transformer has shown impressive success across the natural language processing and visual-related tasks, credit to the long-range modeling capability. In this paper, we leverage the multi-head attention mechanism to update support feature embeddings and query feature embedding with the co-adapting scheme.

In particular, the input to the TEL module consists of a Query feature vector x_q, a Key feature vector x_k and a Value vector x_v. We first map the x_q, x_k and x_v into matrices Q, K, and V, respectively. Then, the attention layer can be formulated as:

$$\hat{A}(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V, \tag{4}$$

where $[Q; K; V] = [W_q x_q; W_k x_k; W_v x_v]$, in which $W_k, W_q, W_v \in \mathbb{R}^{d\times d}$ are parameter matrices, d is the dimension of the input feature vector and we assume all the inputs have the same dimension d by default. The attention layer computes the attention weights by first matching Query elements against all Key elements, where each Key element has a Value element, and then aggregating elements from the Value feature vector through matrix multiplication. In the few-shot action recognition setting, we have $x_q = x_k = x_v = X$, where X are the normalized features obtained from the SGIN module.

Multi-head attention layer performs multiple attention operations on h subspaces and then we concatenate the results together:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h), \tag{5}$$

where $head_i = \hat{A}(Q_i, K_i, V_i)$, and the input $[Q_i, K_i, V_i]$ denotes the i-th subspace from $[Q, K, V]$ with the dimension of d/h. In this work, we employ $h = 8$ heads and for each head we use $d/h = 64$. The Transformer-based embedding learning module receives the support and query features as bags without orders and contextualizes over them, enabling strong co-adaptation between them and generating refined feature embeddings R that are permutation-invariant and task-specific.

Visual-semantic similarity. The feature embeddings R obtained from the TEL module can be regarded as the class prototypes. Specifically, we denote $\mathcal{S}_\alpha$ as the subset of the support set $\mathcal{S}$, which contains videos belonging to class α. Following the classical ProtoNet (Snell et al. (2017)), the prototype $\mathcal{P}_\alpha$ can be calculated by averaging all support feature embeddings in $\mathcal{S}_\alpha$, which can be described as follows:

$$\mathcal{P}_\alpha = \frac{1}{|\mathcal{S}_\alpha|} \sum_{x_i \in \mathcal{S}_\alpha} R_i, \tag{6}$$

where $\mathcal{S}_\alpha = \{(x_i, y_i)|x_i \in \mathcal{S}, y_i = \alpha\}$, R_i is the feature embedding of video x_i. Then, we can calculate the distance similarity between the query embedding q and the support prototype of class α:

$$\mathcal{L}_p^\alpha = \|R_q - \mathcal{P}_\alpha\|, \tag{7}$$

where $\|\cdot\|$ denotes the distance measurement function. Different from ProtoNet (Snell et al. (2017)), we choose the cosine similarity to compute the probabilities between the support set and query set rather than using the Euclidean distance.

Instead of directly applying a softmax layer over the distance similarities to get the final prediction, we propose to incorporate the semantic information of the training class labels to further discriminate the class-similarity among the support-query pairs and boost the performance of the proposed EANT. Specifically, we first use Word2Vec (Mikolov et al. (2013)) to extract semantic embeddings for all the sampled training class labels. Then, we employ a two-layer multi-layer perception (MLP) with LeakyReLU activation, which maps the word embeddings in semantic space to the visual space, increasing the representation ability of the embeddings. After that, we compute the cosine similarity scores between the query embedding and all the transformed semantic embeddings. Finally, we combine the visual-semantic pair similarity scores with the $\mathcal{L}_p^\alpha$ and use a softmax layer to obtain the final predicted result of the query sample q that belongs to class α:

$$\mathcal{G}_\theta(y = \alpha|q) = \frac{exp(\mathcal{L}_p^\alpha + \lambda \|R_q - V_\alpha\|)}{\sum_{i=1}^{N} exp(\mathcal{L}_p^i + \lambda \|R_q - V_i\|)}, \tag{8}$$

where $\lambda \geq 0$ is a hyper-parameter to control the contribution of the visual-semantic similarity. V_α denotes the transformed semantic embedding of class α. y is the predicted label for the query sample q. In particular, for $\lambda = 0$, it becomes the original ProtoNet model.

4. Experiments

In this section, we first describe the datasets and the implementation details of the proposed method. Then, we perform comprehensive experiments to demonstrate its effectiveness on different few-shot action recognition datasets and compare its performance with other state-of-the-art methods. Finally, we provide an ablation study of the proposed method and visualize some examples to intuitively analyze the impact of each component.

4.1. Experimental Setup

Datasets. To evaluate the performance of our proposed method, we conduct extensive experiments on three challenging few-shot action recognition datasets: Kinetics (Kay et al. (2017)), UCF101 (Soomro et al. (2012)) and HMDB51 (Kuehne et al. (2011)). Kinetics is a large-scale dataset with 400 action classes and it contains 306,245 videos. We follow the split strategy proposed by CMN (Zhu and Yang (2018)), which samples 100 action classes from the original 400 action classes, adopting 64, 12, 24 non-overlapping action classes as the training set, validation set, and testing set, respectively. UCF101 consists of 101 action classes with 13,320 videos. We follow the split setting used in Zhang et al. (2020), where 70, 10, 21 disjoint action classes are sampled as the training set, validation set, and testing

set, respectively. HMDB51 has 51 action classes with 6,766 videos. We follow the same split protocol introduced in Zhang et al. (2020), in which 31, 10, 10 disjoint action classes are selected as the training set, validation set, and testing set, respectively.

Implementation details. We follow the sparse sampling strategy described in TSN (Wang et al. (2016)), which divides each input video into $T = 8$ segments and then randomly samples one frame in each segment. For each frame, we first resize it to 256×256 and then apply a random crop with the size of 224×224. We use ResNet-50 (He et al. (2016)) pre-trained on ImageNet (Deng et al. (2009)) as the backbone. We first finetune it on the training data for 8 epochs, where the learning rate is set to 1×10^{-4}. Then, we finetune our EANT for another 8 epochs in a meta-learning way but remain the backbone unchanged. Each epoch contains a total of 2000 training episodes with the learning rate of 2×10^{-5}. We utilize the stochastic gradient descent (SGD) with momentum=0.9 to optimize our EANT parameters. For the UCF101 and HMDB51 datasets, we set the learning rate to 1×10^{-5}. We train our EANT under the 5-way 1-shot setting and evaluate it on the 5-way K-shot benchmark, where $K = 1, 2, 3, 4, 5$. During the meta-testing phase, the mean accuracy over 10,000 random test episodes is reported.

4.2. Comparison with State-of-the-art Methods

Baseline. We adopt ResNet-50 pre-trained on ImageNet as our backbone to extract visual features and use ProtoNet (Snell et al. (2017)) to obtain the prototypes for each class in the support set. We choose the cosine similarity rather than the Euclidean distance used in ProtoNet for the similarity computation between the support set and the query set. Besides, we utilize a channel-wise 1D convolution to obtain the temporal information.

Results on Kinetics. We compare the proposed EANT with several state-of-the-art few-shot action recognition methods on the Kinetics dataset. The comparison results are shown in Table 1. Our EANT significantly outperforms all the competitors and the baseline in all shot settings. We achieve new state-of-the-art results with 75.3%, 82.6%, 85.1%, 86.4%, and 87.5% under the 1-shot, 2-shot, 3-shot, 4-shot, and 5-shot settings, respectively. We notice that compared with the baseline, our EANT achieves an improved performance of 75.3%, with a significant absolute gain of 8.4% under the 1-shot setting. In addition, our EANT performs favorably against AMeFu-Net (Fu et al. (2020)) by achieving gains of 1.2%, 1.5%, 0.8%, 0.8%, and 0.7% under the 1-shot, 2-shot, 3-shot, 4-shot, and 5-shot settings, respectively. Note that AMeFu-Net takes 16 frames as input while we only utilize 8 frames as input. The superior results verify that our EANT can achieve excellent performance even with fewer input frames. Compared with current state-of-the-art methods, such as ITANet (Zhang et al. (2021)), TRX (Perrett et al. (2021)), and HyRSM (Wang et al. (2022)), our EANT outperforms these methods by 1.7%, 11.7%, and 1.6% under the 1-shot setting. The experimental results demonstrate that our embedding adaptation network with Transformer and incorporating semantic knowledge can achieve superior performance.

Results on UCF101 and HMDB51. We also evaluate the proposed EANT on the UCF101 and HMDB51 datasets. We compare our EANT with several state-of-the-art methods under the 1-shot, 3-shot, and 5-shot settings, and the results are listed in Table 2. We observe that our EANT achieves a new state-of-the-art on both datasets. Concretely, on

Table 1: Comparison with other state-of-the-art methods on the Kinetics dataset. We report 5-way few-shot action recognition accuracy (%) on the meta-testing set. "-" denotes the result is not available in published works.

Methods	1-shot	2-shot	3-shot	4-shot	5-shot
Baseline	66.9	77.7	80.5	81.6	83.6
Matching Net (Vinyals et al. (2016))	53.3	64.3	69.2	71.8	74.6
MAML (Finn et al. (2017))	54.2	-	-	-	75.3
CMN (Zhu and Yang (2018))	60.5	70.0	75.6	77.3	78.9
TARN (Bishay et al. (2019))	66.6	74.6	77.3	78.9	80.7
Embodied Learning (Fu et al. (2019))	67.8	77.8	81.1	82.6	85.0
ARN (Zhang et al. (2020))	63.7	-	-	-	82.4
TAM (Cao et al. (2020))	73.0	-	-	-	85.8
AMeFu-Net (Fu et al. (2020))	74.1	81.1	84.3	85.6	86.8
ITANet (Zhang et al. (2021))	73.6	-	-	-	84.3
TRX (Perrett et al. (2021))	63.6	76.2	81.8	83.4	85.9
HyRSM (Wang et al. (2022))	73.7	80.0	83.5	84.6	86.1
EANT	**75.3**	**82.6**	**85.1**	**86.4**	**87.5**

the UCF101 dataset, we achieve 87.0%, 94.6%, and 96.2% under the 1-shot, 3-shot, and 5-shot settings, outperforming AMeFu-Net (Fu et al. (2020)) by 1.9%, 1.5%, and 0.7%, respectively. Besides, our EANT outperforms HyRSM (Wang et al. (2022)) by 3.1%, 1.6%, and 1.5% under the 1-shot, 3-shot, and 5-shot settings, respectively. On the HMDB51 dataset, compared to AMeFu-Net, we achieve gains of 2.3%, 1.6%, and 1.7% under the 1-shot, 3-shot, and 5-shot settings, respectively. And our EANT outperforms HyRSM by 1.2%, 1.4%, and 1.2% under the 1-shot, 3-shot, and 5-shot settings, respectively. The consistent improvements on both datasets demonstrate the advanced generalization ability of the proposed EANT for different scenes.

4.3. Ablation Studies

To analyze the impacts of different components in the proposed EANT, we conduct ablation studies on the Kinetics dataset. Generally, we perform ablative experiments on the 5-way setting and summarize the results of the 1-shot, 3-shot, and 5-shot settings in Table 3.

Impact of the self-guided instance normalization module. We first analyze the impact of the self-guided instance normalization (SGIN) module. As shown in Table 3, we observe that the SGIN module is highly effective. Specifically, compared to the method (a) representing the baseline, the method (b) using the SGIN module brings 7.6%, 3.0%, and 2.2% performance gains under the 1-shot, 3-shot, and 5-shot settings, respectively. In particular, we notice that the performance gains under the 1-shot and 3-shot settings are remarkable, which indicates that the SGIN module can help to learn rich and effective

Table 2: Comparison with other state-of-the-art methods on the UCF101 and HMDB51 datasets. We report 5-way few-shot action recognition accuracy (%) on the meta-testing set. "-" denotes the result is not available in published works.

Method	UCF101			HMDB51		
	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
Baseline	78.7	90.3	92.9	49.5	62.9	68.0
ProtoGAN (Dwivedi et al. (2019))	62.3	75.6	80.5	35.7	46.6	51.5
ARN (Zhang et al. (2020))	66.3	-	83.1	45.5	-	60.6
AMeFu-Net (Fu et al. (2020))	85.1	93.1	95.5	60.2	71.5	75.5
TRX(Perrett et al. (2021))	78.2	92.4	96.1	53.1	66.8	75.6
HyRSM (Wang et al. (2022))	83.9	93.0	94.7	60.3	71.7	76.0
EANT	**87.0**	**94.6**	**96.2**	**62.5**	**73.1**	**77.2**

Table 3: Ablation studies on the self-guided instance normalization module (SGIN), the Transformer-based embedding learning (TEL) module and the visual-semantic similarity (VSS). We report 5-way few-shot action recognition accuracy (%) on the Kinetics dataset.

Methods	Baseline	SGIN	TEL	VSS	1-shot	3-shot	5-shot
(a)	✓				66.9	80.5	83.6
(b)	✓	✓			74.5	83.5	85.8
(c)	✓		✓		74.3	83.5	85.7
(d)	✓			✓	72.8	82.4	84.8
(e)	✓	✓	✓	✓	**75.3**	**85.1**	**87.5**

feature embeddings with few samples. The superior performance verifies that the improved SGIN module benefits the proposed EANT to achieve higher recognition accuracy.

Impact of the Transformer-based embedding learning module. In addition, we assess the impact of the Transformer-based embedding learning (TEL) module. As can be seen in Table 3, when using the TEL module, the accuracy of the method (c) reaches 74.3%, which is 7.4% higher than the method (a) under the 1-shot setting. The experimental results demonstrate that adapting the feature embeddings derived from the support samples and the query samples to the target classification task is effective for few-shot action recognition. The TEL module can contextualize over the input samples, enabling strong co-adaptation of each sample in an episode.

Impact of the visual-semantic similarity. We also analyze the impact of the visual-semantic similarity between the query visual embedding and the semantic embeddings of the training class labels. As shown in Table 3, we can see that compared to the method

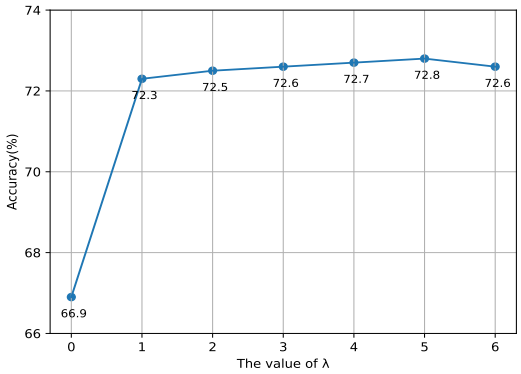


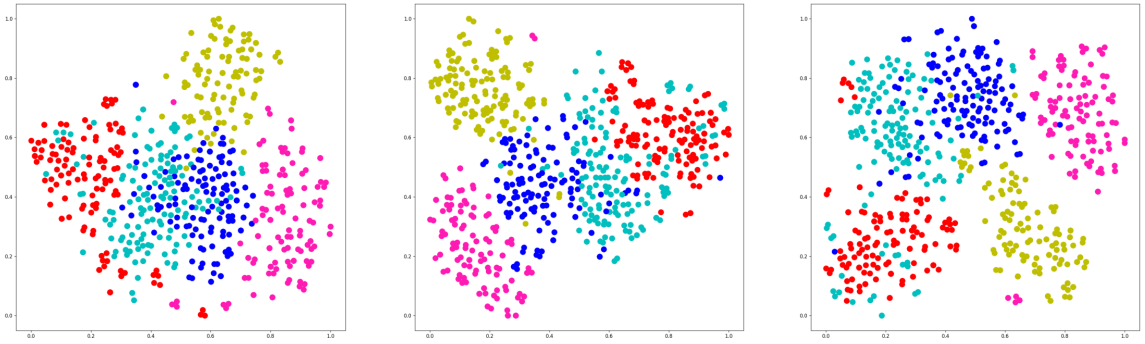
Figure 2: Performance on Kinetics with different values of the hyper-parameter λ under the 5-way 1-shot setting.

(a), the method (d) brings 5.9% performance gain under the 1-shot setting. Moreover, the method (e) establishs a new state-of-the-art on the Kinetics dataset, which boosts the recognition accuracy to 75.3%, 85.1%, and 87.5% under the 1-shot, 3-shot, and 5-shot settings, respectively. The experimental results verify that stacking the SGIN module with the TEL module and incorporating the semantic knowledge of the training class labels can definitely improve the performance of the proposed EANT.

Impact of the hyper-parameter λ. We conduct experiments to explore the impact of different values of the hyper-parameter λ on performance. λ is used to control the contribution of the visual-semantic similarity. We set the value of λ from 0 to 6. The results are shown in Figure 2. We can observe that as the value of λ increases, the performance improves. When setting $\lambda = 5$, the network can achieve the best performance of 72.8% under the 1-shot setting. Thus, we set $\lambda = 5$ for all the experiments.

4.4. Visualization Results

To qualitatively show the discriminative capability of the class-specific feature embeddings learned by the SGIN module and the superior performance of the proposed EANT, we visualize five support classes with and without the SGIN module under the 5-way 1-shot setting in Figure 3. We employ the t-SNE visualization (van der Maaten and Hinton (2008)) to map the video-level features of each video in $\mathcal{D}_{test}$ into a two-dimension metric space. We can intuitively see that by combining the baseline with the SGIN module, each cluster appears more concentrated and the distance between inter-class increases. This demonstrates that the SGIN module can enable the network to yield more class-specific and discriminative feature embeddings, which are good for few-shot action recognition. Besides, we quantitatively observe that the feature embeddings learned by the proposed EANT are more semantically separable compared to the baseline and the baseline with the SGIN module, demonstrating the strong performance of the proposed EANT.



(a) Baseline (b) Baseline+SGIN (c) EANT

Figure 3: t-SNE visualization of support samples distribution in a 5-way 1-shot task on the HMDB51 dataset. (a) corresponds to the result of the baseline. (b) shows the result of the baseline with the SGIN module. (c) represents the result of the proposed EANT. Different colors represent different classes.

5. Conclusions

In this paper, we present a novel embedding adaptation network with Transformer (EANT) for few-shot action recognition. We first utilize a new self-guided instance normalization module to improve the discriminative feature representations learning ability of the proposed EANT, generating class-specific feature embeddings. Then, we design a Transformer-based embedding learning module to customize the feature embeddings by jointly considering the vital contextual cues cross different videos in an episodic task to obtain task-specific feature embeddings. We further propose to incorporate the semantic information of the training class labels, leveraging it as an additional supervisory signal to help the proposed EANT for obtaining more roubst and discriminative feature representations. Experiments on three challenging few-shot action recognition datasets validate the effectiveness of our method.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants U21A20514, 62122011, 61872307 and 62002302; in part by the China Fundamental Research Funds for the Central Universities under Grant 20720210099.

References

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021.

Mina Bishay, Georgios Zoumpourlis, and I. Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. In *BMCV*, 2019.

Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *CVPR*, 2020.

Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. Spact: Self-supervised privacy preservation for action recognition. In *CVPR*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Sai Kumar Dwivedi, Vikram Gupta, Rahul Mitra, Shuaib Ahmed, and Arjun Jain. Protogan: Towards few shot learning for action recognition. In *ICCV Workshops*, 2019.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

Yuqian Fu, Chengrong Wang, Yanwei Fu, Yu-Xiong Wang, Cong Bai, Xiangyang Xue, and Yu-Gang Jiang. Embodied one-shot video recognition: Learning from actions of a virtual embodied agent. In *ACM MM*, 2019.

Yuqian Fu, Li Zhang, Junke Wang, Yanwei Fu, and Yu-Gang Jiang. Depth guided adaptive meta-fusion network for few-shot video recognition. In *ACM MM*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, and et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Hildegard. Kuehne, Hueihan. Jhuang, Estíbaliz. Garrote, Tomaso. Poggio, and Thomas. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *CVPR*, 2021.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Flood Sung, Li Zhang, Tao Xiang, Timothy Hospedales, and Yongxin Yang. Learning to learn: Meta-critic networks for sample efficient learning. *arXiv preprint arXiv:1706.09529*, 2017.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NeurIPS*, 2016.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Val Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

Xiang Wang, Shiwei Zhang, Zhiwu Qin, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hybrid relation guided set matching for few-shot action recognition. In *CVPR*, 2022.

Wenhao Wu, Dongliang He, Tianwei Lin, Fu Li, Chuang Gan, and Errui Ding. Mvfnet: Multi-view fusion network for efficient video recognition. In *AAAI*, 2021.

Bin Xiao, Chien-Liang Liu, and Wen-Hoar Hsaio. Proxy network for few shot learning. In *ACML*, 2020.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018.

Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O. Pinheiro. Adaptive cross-modal few-shot learning. In *NeurIPS*, 2019.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. In *NeurIPS*, 2019.

Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *CVPR*, 2020.

Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip H. S. Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *ECCV*, 2020.

Songyang Zhang, Jiale Zhou, and Xuming He. Learning implicit temporal alignment for few-shot video classification. In *IJCAI*, 2021.

Yuanding Zhou, Baopu Li, Zhihui Wang, and Haojie Li. Video action recognition with neural architecture search. In *ACML*, 2021.

Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *ECCV*, 2018.