3D-aware Select, Expand, and Squeeze Token for Aerial Action Recognition

Luying Peng , Xiangbo Shu*, Yazhou Yao, Guosen Xie

Nanjing University of Science and Technology
pengluying@njust.edu.cn, shuxb@njust.edu.cn, yazhou.yao@njust.edu.cn, guosen.xie@njust.edu.cn

Abstract

Aerial Action Recognition (AAR) in videos captured by Unmanned Aerial Vehicles (UAVs) plays a vital role in numerous applications. However, current methods related to traditional action recognition primarily cater to fixed or near cameras, and rarely consider the movement disturbance of UAVs, including their varying attitudes and positions. Those characteristics of aerial videos bring moving objects in small regions compared to broad backgrounds and relative movement to the motion of objects, which reflect more sparse and disturbed semantic information for AAR. To address these issues, we present a novel framework, dubbed 3D-Tok, to Select, Expand, and Squeeze original visual tokens for obtaining compact yet diverse semantic-enhanced tokens. In particular, we present a 3D-token selector (3TS) to select complex yet diverse tokens in three channels, capturing the semantic awareness of moving objects in comparatively small regions. Additionally, to get rid of disturbed semantic information caused by the UAV flight, we present an Expand-Squeeze Converter (ESC) to adaptively expand and squeeze the 3D-selected tokens constrained by contrastive loss, thereby suppressing the semantic-irrelevant information and reinforce semantic-relevant information via the interpolation converting.

Introduction

With the rapid increase of Unmanned Aerial Vehicles (UAVs), aerial videos have drawn ever-growing attention due to their wide range of applications and flexibility in various situations. The increasing applications of UAVs have led to a rising quantity of aerial videos, underscoring the importance of aerial action recognition. Nonetheless, compared to traditional action recognition in videos shot on fixed ground cameras, the human entity in UAV videos appears rather small due to high camera altitudes (Figure 1). In addition, flying UAVs in the air cause different motion durations and diverse perspectives. These factors engender notable obstacles in UAV-based aerial action recognition and call for designing specific recognition approaches that are purposefully crafted to address these specific traits.

Convolutional Neural Networks (CNN) have reached great success in general action recognition tasks (Koohzadi and Charkari 2017). These methods excel at capturing local spatial patterns in data, making them ideal for tasks like recognition, detection, and classification tasks. Whether 2D convolutions (Li et al. 2016) or 3D convolutions (Tran et al. 2015), both are better at capturing spatial features and may fall slightly short in modeling temporal relations. Nonetheless, in UAV-based videos, temporal relations are more complex due to the motion interference brought by the fast motion of flying cameras and broad backgrounds contain repetitive and similar semantic information.

In recent years, transformer-based methods (Arnab et al. 2021; Fan et al. 2021) have been developed for general action recognition tasks due to their exceptional temporal modeling capabilities and powerful self-attention mechanisms. Whereas in UAV videos, moving objects only occupy a small region compared to the background, and flying

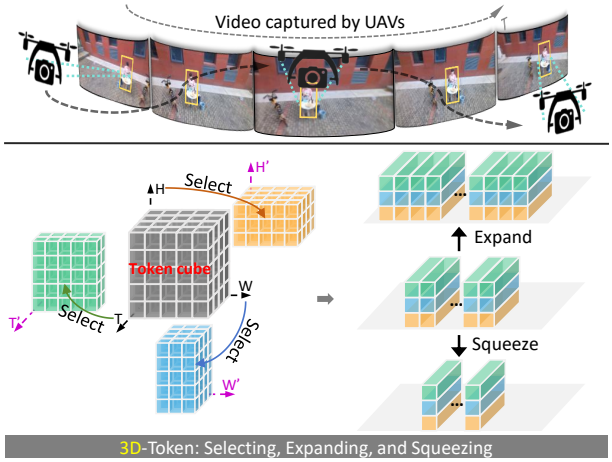

Figure 1: Insights of this work. In videos captured by UAVs, moving objects are in small regions compared to broad backgrounds and the flying UAVs bring the disturbance of relative movement to the motion of objects, which reflects more sparse and disturbed semantic information. Our approach can: i) select compact yet diverse tokens for capturing the semantic awareness of moving objects in comparatively small regions; and ii) further expand and squeeze tokens for reinforcing semantic-relevant information while suppressing semantic-irrelevant information.

*Corresponding author.

UAVs with varying attitudes and positions bring the disturbance of relative movement to the motion of objects. These characteristics of UAV videos make it even more difficult for Transformer to capture sparse and disturbed semantic information, due to the lack of inductive bias usage in Transformer (Shi et al. 2023). Therefore, a tailor-made feature refinement mechanism is needed to filter out irrelevant information, as relying solely on the transformer's global attention mechanism may not be sufficient to address this challenge.

Based on the above analysis, we develop a novel framework for aerial action recognition, called 3D-Tok, which comprehensively considers small moving objects, as well as the disturbance of relative movement performed by UAVs. The overall framework of 3d-Tok is composed of two feasible modules, namely 3D-Token Selector (3TS) and Expand-squeeze Converter (ESC), as shown in Figure 2. Specifically, 3TS aims to select complex yet diverse 3D tokens along three channels, which discards the redundant unimportant tokens in broad backgrounds to eliminate the interference of cluttered information. ESC aims to leverage semantic-level expand-squeeze interpolation to further concentrate and enrich the semantic information of 3D-selected tokens. By involving the selecting, expanding, and squeezing operations of tokens into an all-in-one framework, the proposed 3D-Tok effectively acquires sufficient semantic information concerning spatial-temporal dynamics for the AAR task.

Overall, the main contributions of this work are threefold:

- **New solution for AAR task.** We propose a novel 3D-Token selecting transformer with the expand-squeeze converter (3D-Tok) to address the problem of aerial action recognition by selecting compact yet diverse tokens in a multiple channel-wise way and further enhancing their semantic information in the expand-squeeze interpolation way.
- **New token selecting strategy.** A new 3D-Token Selector (3TS) is presented to synchronously select compact yet diverse tokens along with the three channels via selection networks, capturing the semantic awareness in comparatively small regions captured by varying UAVs with different attitudes, positions, and views.
- **New semantic enhancing strategy.** A new Expand-Squeeze Converter (ESC) is presented to reinforce semantic-relevant information via expand-interpolation converting while suppressing the semantic-irrelevant information via squeeze-interpolation converting, further enhancing the semantic awareness of 3D-selected tokens.

Related Works

Video Understanding

Video understanding tasks aim to recognize and understand content from videos automatically. It has gained widespread attention and applied to various tasks, such as motion prediction (Shu et al. 2021) and action recognition (Cao et al. 2023; Shu et al. 2022). In motion prediction, it focuses on predicting the motion trajectories of actors within a video. It focuses on predicting the motion trajectories of actors within the video. Earlier methods are mainly categorized into physics-based motion prediction models and interaction-aware motion prediction models (Dagli and Reichardt 2002). For action recognition, many traditional approaches relied on the design of hand-crafted features, such as (Wang and Schmid 2013). Later, a wide range of CNN-based approaches are proposed, including two-stream frameworks (Lin, Gan, and Han 2019) and 3D CNNs (Tran et al. 2015). More recently, existing methods have expanded from CNNs to Transformers (Arnab et al. 2021) based on attention mechanisms. However, the above methods are commonly used for general video understanding captured by ground-based cameras, and cannot handle UAV-based video understanding tasks well due to the small moving objects and unfixed cameras.

Aerial Action Recognition

Aerial action recognition is a subfield of action recognition, characterized by the identification of human actions in videos captured by UAVs. Early studies utilize pose-based methodologies to extract local representation within a global receptive field to understand actions in aerial videos. Building on the progress in Deep Neural Networks (DNN), Some methods treat aerial videos as a collection of frames, using 2D-CNN (Geraldes et al. 2019) to extract action features at every frame, and then integrate features to identify the actions within the video. Other methods employ the I3D network (Joao and Andrew 2017) to capture spatial-temporal features in aerial videos directly. With the widespread adoption of attention mechanisms, Kothandaraman et al. (Kothandaraman et al. 2022) introduced an attention mechanism based on the Fourier Transform to focus motion salience. MG Sampler (Zhi et al. 2021) and PMI Sampler (Xian et al. 2024) employed frame selection techniques to acquire semantic-compact frames for learning the high-purity discriminative features, and achieve the satisfied performance. These frame selection techniques can be seen as the selection on the temporal dimension. Unlike these techniques, we propose to select tokens in three dimensions by an all-in-one framework, covering both spatial and temporal dimensions.

Token Selection Mechanism

The token selection mechanism is first explored by (Zhao et al. 2019), facilitating adaptive and efficient attention modeling in NLP tasks. Some works (e.g. (Wang et al. 2022b)) suggest employing k-nearest Neighbors (k-NN) attention mechanisms for selecting the top-k tokens with the highest similarity. Specifically, they compute the pixel-wise similarity of query-key pairs and perform spatial-wise selection. In the video understanding area, similar patch-wise masking strategies have been investigated for boosting vision transformers (He et al. 2022; Wang et al. 2022a). The patch-wise strategy focuses on local pixel similarity, which may lead to the neglect of global contextual information. Unlike the above methods, we propose to calculate the semantic important score of each token slice to select the tokens with the highest scores in the overall context, maintaining the critical global information while processing more detailed parts.

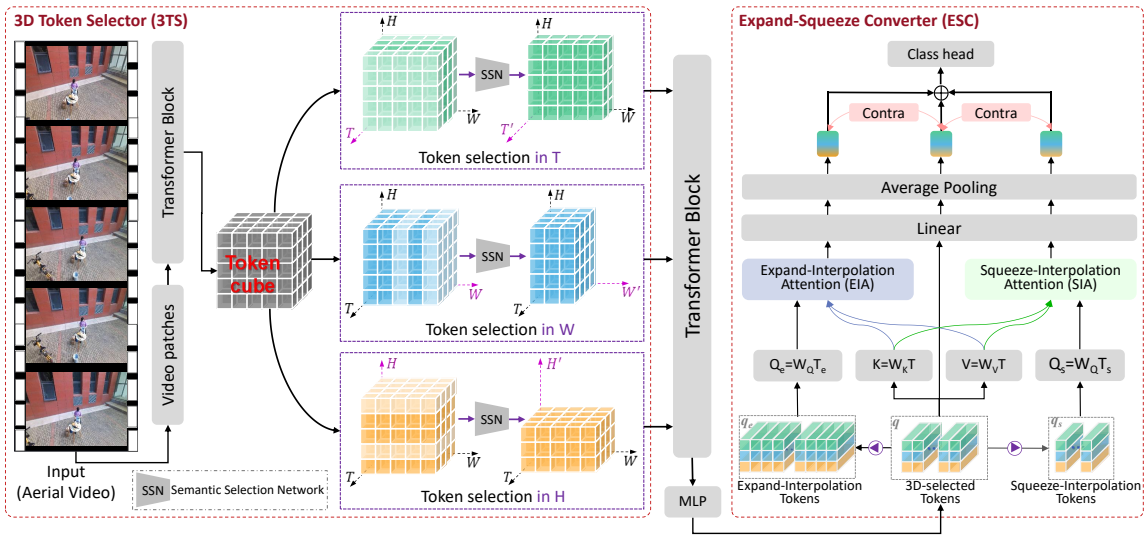

Figure 2: Overall framework of the proposed 3D-Tok. It mainly consists of the 3D-Token Selection module (3TS), and the Expand-Squeeze Conversion module (ESC). Aerial Video frames are firstly patched to 3D embeddings and tokenized to a sequence of tokens in the transformer block. 3TS is utilized to select the three types of compact tokens along three channels, respectively. This makes the tokens compact yet diverse. Following 3TS, ESC is utilized to expand and squeeze 3D-selected tokens via semantic–level interpolation constrained by the contrastive loss. This further reinforces semantic-relevant information while suppressing the semantic-irrelevant information.

Methodology

Overview

The overall framework of the proposed 3D-Tok is shown in Figure 2, which is mainly composed of 3TS and ESC. For the input of a video $\boldsymbol{X} \in \mathbb{R}^{T\times H\times W}$ consisting of T frames of $H \times W$ resolution, the network firstly patches video frames to 3D embeddings, and utilize a Transformer Block to produce spatiotemporal tokens. In **3TS**, directly extracted a 3D token cube (denoted by $\boldsymbol{q} \in \mathbb{R}^{T\times H\times W\times C}$) from the input videos $\boldsymbol{X}$ are performed selection in three dimensions of T, H, and W, attending to salient frames and delving into the most important Top-K of T, H, and W dimension, which effectively provides more essential representations in UAV videos within sparse semantic motions. After that, the Transformer Block continues modeling three types of selected tokens (denoted by $\boldsymbol{q}^t =\in \mathbb{R}^{T^{'}\times H\times W\times C}$, $\boldsymbol{q}^h =\in \mathbb{R}^{T\times H^{'}\times W\times C}$, and $\boldsymbol{q}^w =\in \mathbb{R}^{T\times H\times W^{'}\times C}$), followed by the MLP layer to recombinate and reshape them for obtaining 3D-selected token $\bar{\boldsymbol{q}} \in \mathbb{R}^{T^{'}\times L^{'}\times C}$. Here, $T^{'}$ denotes one dimension size that will be expanded/squeezed in the following step, and $L^{'}$ means the reshaped combination by the other dimension size. In **ESC**, the 3D-selected token is implemented in the expand interpolation and squeeze interpolation along with the scale Z_e and Z_s, respectively. This process strengthens the semantic-relevant information, meanwhile diminishing the semantic-irrelevant information through converting. To accommodate tokens at different scales, we design the corresponding adaptive attention mechanism—Expand Interpolation Attention and Squeeze Interpolation Attention, respectively modeling $\boldsymbol{q_e}$ and $\boldsymbol{q_s}$. Finally, the obtained action class tokens are summed of the normal, expanded, and squeezed tokens, of which consistency is controlled by contrastive loss.

3D-Token Selector (3TS)

Considering the small moving objects in the broad background captured by the unsteady flying UAVs, considerable significance to spatial information with temporal motions is assigned. To alleviate the redundant patches and unneeded temporal clues brought by blank backgrounds with aerial vehicles, we present A new 3D-Token Selector (3TS) to synchronously select compact and diverse tokens along with the three channels via semantic selection networks using the Top-K policy.

Semantic Selection Network. Given a sequence of input token cube $\boldsymbol{q}^d \in \mathbb{R}^{M\times M_1\times M_2\times C}$, the goal of the semantic important network is to generate a semantic-aware importance score for each token slice along with the dimension M. Here, M denotes the length of one dimension (will be selected), as well as M_1, and M_2 denotes the lengths of the other two dimensions (will be un-selected). Specifically, to integrate semantic-aware important scores of each token slice, we adopt activated linear projection layers. Firstly, we map the input tokens $\boldsymbol{q}^d$ into the representations of the token slices $\{\boldsymbol{x}_m^l\}_{m=1}^M$ to perceive the local relationships via a linear projection, as follows,

$$\{\boldsymbol{x}_m^l\}_{m=1}^M = \mathrm{Linear}(\mathrm{LN}(\boldsymbol{q}^d), \boldsymbol{\omega}_1), \tag{1}$$

where $\mathrm{LN}(\cdot)$ is Layer Normalization. Then, we leverage the global information by computing the average of $\{\boldsymbol{x}_m^l\}_{m=1}^M$

to acquaint a representation $\boldsymbol{x}^g$ as the global view, subsequently concatenate it with $\boldsymbol{x}_m^l$ to generate the norm token sequence, thus effectively capture the important token slice to the overall semantics. In particular, the concatenated global representation can be re-written as:

$$\hat{\boldsymbol{x}} = \{[\boldsymbol{x}_1^l, \boldsymbol{x}^g], \cdots, [\boldsymbol{x}_m^l, \boldsymbol{x}^g], \cdots, [\boldsymbol{x}_M^l, \boldsymbol{x}^g]\}, \quad (2)$$

The concatenated feature $\hat{\boldsymbol{x}}$ is then fed to two layers of activated linear projection to generate the importance scores:

$$\boldsymbol{s} = \mathrm{Act}(\mathrm{Linear}(\hat{\boldsymbol{x}}, \boldsymbol{\omega_2})), \quad (3)$$

where $\mathrm{Act}(\cdot)$ is the GELU activation function. After getting the score vectors of all token slices, we normalized them into $\bar{\boldsymbol{s}}$ with the min-max normalization.

Top-K Policy in 3D Selection. Via the semantic selection network, we generate score vector $\bar{\boldsymbol{s}}$ for selecting K highest scores and extracting the corresponding token slices. Specifically, Using the Top-K operation to get the indicator set $\Omega = \{\mathrm{Top\_K}(\bar{\boldsymbol{s}})\} \in \mathbb{N}^K$ ($K < M$), and them obtain one 3D-selected token $\boldsymbol{q}^t$ based on dimension M, as follows,

$$\boldsymbol{q}^M = \mathrm{Selection}(\boldsymbol{q}^d, \Omega), \quad (4)$$

where $\mathrm{Selection}(\cdot)$ is the selection operator under the indicator set. Based on Eq. (1)-(4), given the 3D token cube $\boldsymbol{q} \in \mathbb{R}^{T \times H \times W \times C}$, we can obtain the three types of two 3D-selected tokens $\boldsymbol{q}^t$, $\boldsymbol{q}^h$ and $\boldsymbol{q}^w$ at three dimensions, respectively. Finally, we obtain the recombined and reshaped 3D-selected token $\bar{\boldsymbol{q}} \in \mathbb{R}^{T^{'} \times L^{'} \times C}$ via Transformer Block and MLP layer.

Expand-Squeeze Converter (ESC)

Given UAV-based videos, capturing motion patterns requires sufficient semantic information. As shown in Figure 1, flying UAVs bring cluttered backgrounds and relative movement. It leads to a challenge for straightly extracting and then effectively modeling features. Thus, we present an Expand-Squeeze Converter (ESC) to adaptively model selected compact yet diverse tokens via Expand/Squeeze Interpolation. Here, to ensure consistency among these normal, expanded and squeezed tokens, they are interacted by Expand/Squeeze Interpolation Attention and then constrained by contrastive Loss. It is noted that the Expand/Squeeze Interpolation can be creatively implemented at any dimension. Although we take the interpolation at one dimension as an example in this paper, the interpolation respectively at all dimensions can be easily implemented.

Expand/Squeeze Interpolation. After obtaining the selected tokens $\bar{\boldsymbol{q}} \in \mathbb{R}^{T^{'} \times L^{'} \times C}$, we present Expand/Squeeze Interpolation (including three component layers of a Batch Normalization, a 3D convolution, and a 3D deconvolution) at dimension T' to reinforce semantic-relevant information via expansion converting, meanwhile suppressing the semantic-irrelevant information via squeeze converting, separately. Particularly, in Expand Interpolation, the expanding scale of tokens is dilated to Z_e in the expanding stream by upscaling tokens along the time dimension using the expanding interpolation operation, which brings in more rich semantic contexts for videos with small spatial/temporal scale. Similarly, in Squeeze Interpolation, we constrict the scale of a series of tokens to Z_s in the squeezing stream along the time dimension using the squeezing interpolation operation, thereby suppressing the excessive information in the big spatial/temporal scale. Formally, the expand-interpolation tokens $\bar{\boldsymbol{q}}_e \in \mathbb{R}^{Z_e \times L^{'} \times C}$ and $\bar{\boldsymbol{q}}_s \in \mathbb{R}^{Z_s \times L^{'} \times C}$ via the expand-squeeze interpolation can be expressed as:

$$\bar{\boldsymbol{q}}_e = \mathrm{Upscale}(\mathrm{BN}(\bar{\boldsymbol{q}})), \quad (5)$$

$$\bar{\boldsymbol{q}}_s = \mathrm{Downscale}(\mathrm{BN}(\bar{\boldsymbol{q}})), \quad (6)$$

where $\mathrm{BN}(\cdot)$ is Batch Normal layer, $\mathrm{Upscale}(\cdot)$ and $\mathrm{Downscale}(\cdot)$ denote the expand interpolation and squeeze interpolations of 3D convolution, respectively.

Expand-Squeeze Interpolation Attention. After Expand/Squeeze Interpolation, the semantic context information is modified by expanding or squeezing token sequences, which may partially generate the weakening semantic information. Following cross-attention, we take $\bar{\boldsymbol{q}}_e$ and $\bar{\boldsymbol{q}}_s$ to interact with the normal tokens $\bar{\boldsymbol{q}}$ for fine complementing semantic information into the expand/squeeze branch via Expand-Interpolation Attention (EIA) and Squeeze-Interpolation Attention (SIA). In particular, we set $\bar{\boldsymbol{q}}_e$ as query $\boldsymbol{Q}_e$, and $\bar{\boldsymbol{q}}$ as key/value ($\boldsymbol{K}$/$\boldsymbol{V}$) for EIA, and obtain the semantic-complemented token sequences $\bar{\boldsymbol{q}}_s$ in the expanding branch, as follows:

$$\boldsymbol{q}_e^{'} = \mathrm{LN}(\mathrm{EIA}(\boldsymbol{Q}_e, \boldsymbol{K}, \boldsymbol{V}) + \bar{\boldsymbol{q}}_e), \quad (7)$$

$$\boldsymbol{F}^e = \mathrm{AP}(\boldsymbol{q}_e^{'} + \mathrm{MLP}(\boldsymbol{q}_e^{'})), \quad (8)$$

where $\mathrm{AP}(\cdot)$ denotes the average pooling, $\mathrm{EIA}(\cdot)$ means the expand interpolation attention, and $\mathrm{MLP}(\cdot)$ represents the MLP layer. Following EIA, the semantic details initially present in the normal tokens have been integrated into $\boldsymbol{F}^e$, preventing any loss or bias of semantic richness. Similar to EIA, we can use SIA to get the $\boldsymbol{F}^s$. Besides, we arrange the normal token representation $\boldsymbol{F}^n$ with the linear layer and the average pooling. Finally, the aggregated action representation $\boldsymbol{F}$ can be obtained through the summary:

$$\boldsymbol{F} = \boldsymbol{F}^e + \boldsymbol{F}^s + \boldsymbol{F}^n, \quad (9)$$

Overall Loss. After obtaining the action representation $\boldsymbol{F}$, we use the cross-entropy loss for the aggregated action representation and the loss can be formulated as:

$$\mathcal{L}_{cls} = -\sum_{i \in \Psi} \boldsymbol{y}_i \log(\hat{\boldsymbol{y}}_i), \quad (10)$$

where Ψ denotes the indicator set of training samples in one batch, $\hat{\boldsymbol{y}}_i = \mathrm{Softmax}(\boldsymbol{F}_i)$, $\boldsymbol{F}_i$ denotes the action representation $\boldsymbol{F}_i$ of the i-th sample calculated by Eq. (9). To ensure the semantic consistency between $\boldsymbol{F}_i$ and the corresponding $\boldsymbol{F}_i^e$/$\boldsymbol{F}_i^s$, we utilize the supervised contrastive loss (Khosla et al. 2020) to train the whole model, as follows:

$$\mathcal{L}_e = \sum_{i \in \Psi} \left(\frac{-1}{|\Psi_2|} \sum_{j \in \Psi_2} \log \frac{\exp(\boldsymbol{F}_i \cdot \boldsymbol{F}_j^e / \tau)}{\sum_{k \in \Psi_1} \exp(\boldsymbol{F}_i \cdot \boldsymbol{F}_k^e / \tau)} \right), \quad (11)$$

$$\mathcal{L}_s = \sum_{i \in \Psi} \left(\frac{-1}{|\Psi_2|} \sum_{j \in \Psi_2} \log \frac{\exp(\boldsymbol{F}_i \cdot \boldsymbol{F}_j^s / \tau)}{\sum_{k \in \Psi_1} \exp(\boldsymbol{F}_i \cdot \boldsymbol{F}_k^s / \tau)} \right), \tag{12}$$

where Ψ_1 is the indicator set of all samples excluding the i-th sample in one batch, $\Psi_2 = \{j \in \Psi_1 : \hat{\boldsymbol{y}}_j^e = \hat{\boldsymbol{y}}_i\}$ is the indicator set of all positive samples, $\hat{\boldsymbol{y}}_j^e = \mathrm{Softmax}(\boldsymbol{F}_j^e)$, τ is a temperature parameter, and the symbol $\cdot$ denotes the dot product. Therefore, by combining with the $\mathcal{L}_{cls}$ in Eq. 10, the overall loss can be described as,

$$\mathcal{L} = \mathcal{L}_{cls} + \boldsymbol{\alpha}\mathcal{L}_e + \boldsymbol{\beta}\mathcal{L}_s, \tag{13}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the hyper-parameters.

Experiments

Datasets

UAV-Human is the largest UAV-based human behavior understanding dataset (Li et al. 2021). This dataset contains 20,728 high-definition videos captured in various indoor and outdoor settings, encompassing a broad range of lighting and weather conditions. The videos cover dynamic backgrounds and UAVs with diverse motions and altitudes, making this dataset highly challenging. A total of 155 unique actions have been annotated, with some being difficult to differentiate, such as squeeze and yawn actions. Compared with existing works, we use split 1 which contains 15,172 and 5,556 videos for training and testing, respectively.

Drone-Action is a dataset for human action classification in aerial videos (Perera, Law, and Chahl 2019), which contains 240 aerial videos across 13 different human actions performed by 10 human actors. Drone-Action is an outdoor video dataset that was captured using a free-flying UAV, with 168 training clips and 72 testing clips.

RoCoG-v2 is a dataset that contains real and synthetic videos from air and ground perspectives (Reddy et al. 2022). We use 87 long real videos captured from the air with 7 action categories for training and 91 rest for testing.

Experimental Settings

We uniformly sample 16/8 frames to generate each input video $\boldsymbol{X} \in \mathbb{R}^{16/8 \times 224 \times 224}$ and apply the standard RandomResizedCrop and RandomHorizontalFlip augmentation strategy for data preprocessing, following (Herzig et al. 2022; Patrick et al. 2021). We use a batch size of 4, and an SGD optimizer using a momentum of 0.9 and a weight decay of 5×10^{-4}. The maximum training epoch is set to 100, with the initial learning rate 10^{-5}, decreased by a factor of 10 for every 20 epochs. τ in Eq. (10) and Eq. (11) is set to 0.1. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in Eq. (12) are set to 0.2 and 0.1, respectively. The whole model is implemented using Pytorch on two NVIDIA RTX 3090 GPUs.

Comparison with State-of-the-art

Results on UAV-Human. Table 1 shows the performance comparison on the UAV-Human Dataset. 3D-Tok achieves the best performance, gaining the performance improvement of 7.7%, 8.1%, and 9.5% under various settings for the number of frames and frame sizes. Although some mainstream action recognition methods, e.g., I3D-M (Joao and Andrew 2017), X3D-M (Feichtenhofer 2020), TimeSformer (Bertasius, Wang, and Torresani 2021), etc, demonstrate the effectiveness of general action recognition tasks, they fail to produce competitive results on the AAR task. For example, as one of the most representative methods for general action recognition, TimeSformer (Bertasius, Wang, and Torresani 2021) only achieves the recognition accuracy of 31.9%. This is much lower than expected because these methods cannot suit UAV-based video understanding well due to the small moving objects and the disturbance of relative movements performed by UAVs. For the AAR-based methods, namely FAR (Kothandaraman et al. 2022) using self-attention in the frequency domain, and ASAT (Shi et al. 2023) modeling dependencies among local patches, they are superior to normal action recognition methods. Moreover, as the current SOTA methods, MG Sampler (Zhi et al. 2021) and PMI Sampler (Xian et al. 2024) also adopting the frame/feature selection strategy achieve accuracies of 53.8%, and 55.0%, respectively. This well demonstrates that the frame/feature selection strategy is mainstream for capturing sparse and disturbed semantic information in the AAR task.

Results on Drone-Action. We also evaluate 3D-Tok compared with the competitive methods on the Drone-Action dataset, as shown In Table 2. Our 3D-Tok continues to perform better than the alternatives, gaining an improvement of 5.6% over the SOTA method (i.e., ASAT (Shi et al. 2023)). Similar to the results reported in Table 1, as the representative action recognition methods, SlowFast (Feichtenhofer et al. 2019) and Video Swin (Liu et al. 2022) respectively perform 86.7% and 90.7%, which are not unacceptable compared with those of AAR-based method, e.g., ASAT (Shi et al. 2023) and FAR (Shi et al. 2023).

Results on RoCoG-v2. We finally conduct the comparison experiment on the RoCoG-v2 dataset, as shown in Table 3. Unsurprisingly, 3D-Tok achieves the best performance, gaining significant improvements of 19.1% and 25.3% under different settings for the number of frames. Specifically, our method explores the spatial-temporal selection and expanding/squeezing in the 3D space, which is superior to AAR-based AZTR (Wang et al. 2023) exploring the spatial zooming and temporal extracting in the 2D space. It is noted that MoViNet A3 (Kondratyuk et al. 2021) is proposed for general action recognition tasks, and only achieves 29.0% / 34.1% for 8 / 20 frames in each video. It is validated that the idea of 3D-Tok is effective for AAR.

Ablation Study

Effect of Each Module. The framework of the proposed 3D-Tok mainly includes two modules, i.e., 3D-Token Selector (3TS), and Expand-Squeeze Converter (ESC). We conduct the ablation study to validate the superiority of each module in terms of recognition accuracy, as shown in Table 3. Here, the baseline method is the framework of 3D-Tok without 3TS and ESC. We can see that: 1) when our method

Method	Frames	Input Size	Initialization	Acc.(%)
I3D-M (Joao and Andrew 2017) [CVPR2017]	8	540 × 960	Kinetics+ImageNet	21.1
X3D-M (Feichtenhofer 2020) [CVPR2020]	8	540 × 540	Kinetics	36.6
FAR (Kothandaraman et al. 2022) [ECCV2022]	8	540 × 540	Kinetics	38.6
DiffFAR (Kothandaraman, Lin, and Manocha 2022) [Arxiv2022]	8	540 × 540	Kinetics	41.9
PMI Sampler (Xian et al. 2024) [WACV2024]	8	540 × 540	Kinetics	47.7
Ours (3D-Tok)	8	540 × 540	Kinetics	**55.4** (+7.7)
FAR (Kothandaraman et al. 2022) [ECCV2022]	8	620 × 620	Kinetics	39.1
MITFAS (Xian, Wang, and Manocha 2024) [WACV2024]	8	620 × 620	Kinetics	46.6
PMI Sampler (Xian et al. 2024) [WACV2024]	8	620 × 620	Kinetics	52.0
Ours (3D-Tok)	8	620 × 620	Kinetics	**60.1** (+8.1)
X3D-M (Feichtenhofer 2020) [CVPR2020]	16	224 × 224	Kinetics	30.6
FAR (Kothandaraman et al. 2022) [ECCV2022]	16	224 × 224	Kinetics	31.9
TimeSformer (Bertasius, Wang, and Torresani 2021) [ICML2021]	16	224 × 224	Kinetics	33.9
ASAT (Shi et al. 2023) [MM2023]	16	224 × 224	Kinetics	39.7
AZTR (Wang et al. 2023) [ICRA2023]	16	224 × 224	Kinetics	47.4
MITFAS (Xian, Wang, and Manocha 2024) [WACV2024]	16	224 × 224	Kinetics	50.8
MG Sampler (Zhi et al. 2021) [ICCV2021]	16	224 × 224	Kinetics	53.8
PMI Sampler (Xian et al. 2024) [WACV2024]	16	224 × 224	Kinetics	55.0
Ours (3D-Tok)	16	224 × 224	Kinetics	**64.5** (+9.5)

Table 1: Performance comparison on the UAV-Human dataset.

Method	Frames	Input Size	Init.	Acc.(%)
X3D-M	16	224 × 224	K	83.4
I3D	16	224 × 224	K	85.5
FuTH-Net	16	-	K+I	88.4
SlowFast	64	224 × 224	I	86.7
TSM	16	224 × 224	K	90.3
Video Swin	16	224 × 224	K	90.7
FAR	16	224 × 224	K	92.7
ASAT	16	224 × 224	K	93.1
Ours (3D-Tok)	16	224 × 224	K	**98.7** (+5.6)

Table 2: Performance comparison on Drone-Action. "K" and "I" denote the Kinetics and ImageNet datasets, respectively.

Method	Frames	Input Size	Init.	Acc.(%)
MoViNet A3	8	256× 256	Kinetics	29.0
AZTR	8	172 × 172	Kinetics	29.5
Ours (3D-Tok)	8	224× 224	Kinetics	**48.6** (+19.1)
MoViNet A3	20	256× 256	Kinetics	34.1
AZTR	20	172 × 172	Kinetics	40.2
Ours (3D-Tok)	16	224 × 224	Kinetics	**63.7** (+25.3)

Table 3: Performance comparison on RoCoG-V2.

is equipped with either 3TS or ESC, the performance is improved over that of baseline; and 2) when our method is equipped with both 3TS and ESC, the performance is significantly improved over that of baseline. It is concluded that the integration of 3TS and ESC is beneficial to model aerial actions in UAV-based videos.

Effect of Each Loss. We conduct the ablation study to validate each loss in Eq. (13). Table 5 shows the performance of 3D-Tok with one/two contrastive loss, or the cross-entropy loss. We observe that 3D-Tok performs worst when there is only cross-entropy loss. There are different degrees

3TS	ESC	Acc.(%)	
		UAV-Human	Drone-Action
		58.8	90.5
✓		60.2	92.7
	✓	63.9	94.9
✓	✓	**64.5**	**98.7**

Table 4: Ablation study for 3TS and ESC.

$\mathcal{L}_{cls}$	$\mathcal{L}_e$	$\mathcal{L}_s$	Acc.(%)	
			UAV-Human	Drone-Action
✓			63.7	96.5
✓	✓		64.0	97.6
✓		✓	63.9	97.2
✓	✓	✓	**64.5**	**98.7**

Table 5: Ablation study for each loss.

$3K/(T+H+W)$	Acc.(%)	
	UAV-Human	Drone-Action
0.5	62.7	97.5
0.6	63.9	98.1
0.7	64.2	98.5
0.8	**64.5**	**98.7**
0.9	64.3	98.7
1	64.3	98.5

Table 6: Diagnostic study of different values of k in 3TS.

of performance improvement achieved by 3D-Tok when we add either contrastive loss. Finally, when all three losses are added, the performance is the best. It is well illustrated the effectiveness of dual contrastive losses.

Diagnostic Study of K in 3TS. We conduct the diagnostic study to investigate the value of K in 3TS. The setting for the value of K is based on the video size, namely T,

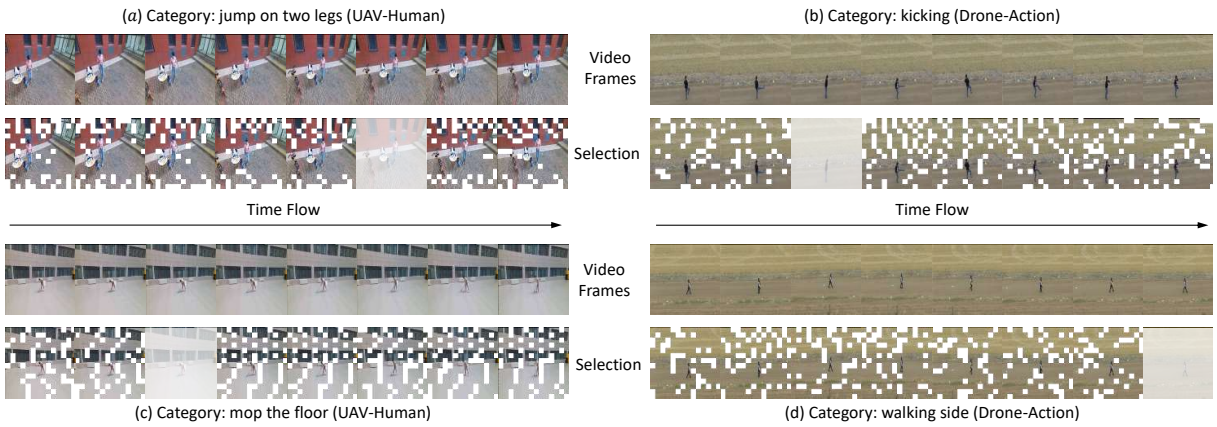

Figure 3: Visualization of 3D-Token selection tokens. After 3D-Tok, some redundant token patches (in the spatial dimension) and token frames (in the temporal dimension) are discarded.

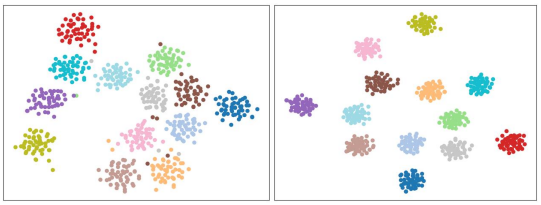

Figure 4: Distribution of action features combined without (left figure) and with (right figure) expanded/squeezed features on UAV-Human. 13 categories are randomly selected for a better view.

W, and H. Thus, a straightforward way is that we set the ratio $r = 3K/(T + W + H) \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ instead of directly setting the value of K. Table 6 shows the performance of 3D-Tok with different values of K. The best performance is achieved when $r = 0.8$. We also find that the performance leads to a dramatic drop when $r = 0.5$ because insufficient spatial/temporal information is available. Noted that utilizing more tokens (e.g., $r = 0.9$) and even all tokens (i.e., $r = 1$) lead to a performance decrease due to interference from cluttered information.

Qualitative Analysis

Visualization of 3D-selected tokens. We give a visualization of 3D-selected tokens to intuitively illustrate the importance of the 3D-aware token-selecting strategy in 3TS. For a better view, we report the visualization comparison between the original video frames and the selected video frames mapped by 3D-selected tokens, as shown in Figure 3. We can see that original video frames contain redundant spatial information (e.g., background), and repetitive temporal information (e.g., similar frames). The selected video frames based on 3D-selected tokens preserve the most meaningful patches (e.g., the human interaction), while filtering out dull backgrounds (e.g., the wild grassland). This further validates the effectiveness of the 3D-aware token-selecting strategy in 3TS.

Visualization of Semantic Discriminatory. To further qualitatively analyze the feature discrimination after enhancing the semantic information of tokens via ESC. We visualize the features with and without the token expanding/squeezing interpolation via the t-SNE tool, as shown in Figure 4. we can see that the final features combined with the interpolated features show a more compact intra-class distribution and a more discriminative inter-class distribution.

Conclusion and Discussion

Conclusion. In this paper, we propose a novel 3D-Tok method to address the problem of AAR. With a 3D-Token Selector (3TS) to select compact yet diverse tokens along three dimensions, and an Expand-Squeeze Converter (ESC) to reinforce semantic-relevant information while suppressing the semantic-irrelevant information via Exapnd-squeeze interpolation conversion, our approach achieved significant performance on three public datasets.

Discussion. Many token selection methods are efficiency-driven, but our work is performance-driven since the current SOTA performance for AAR tasks is unsatisfactory. For 3D-Tok, our primary goal is to comprehensively process data from a three-dimensional perspective to generate higher-quality tokens. We only present the expanding/squeezing in one dimension (i.e., the temporal dimension) due to the space limitation. As shown in Table 1, changing the frame numbers more affects the recognition performance than altering the input sizes. Obviously, the expanding/squeezing in two or three dimensions is easy to implement in our 3D-Tok framework. Moreover, all our token refinement strategies (selection, expansion, and squeeze) are performed in the 3D space, compared to those of existing methods (e.g., MG Sampler, and PMI Sampler) only in the 2D space, our method does not destroy its natural structure. Though the technical designs may be not optimal enough, this is the first attempt to refine tokens in the 3D space.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (Grant No. 62222207, 62072245, and 62276134), the Natural Science Foundation of Jiangsu Province (Grant No. BK20211520).

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, 4.

Cao, M.; Yan, R.; Shu, X.; Zhang, J.; Wang, J.; and Xie, G.-S. 2023. MUP: Multi-granularity Unified Perception for Panoramic Activity Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7666–7675.

Dagli, I.; and Reichardt, D. 2002. Motivation-based approach to behavior prediction. In *Intelligent Vehicle Symposium.*, volume 1, 227–233.

Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; and Feichtenhofer, C. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6824–6835.

Feichtenhofer, C. 2020. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 203–213.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Geraldes, R.; Goncalves, A.; Lai, T.; Villerabel, M.; Deng, W.; Salta, A.; Nakayama, K.; Matsuo, Y.; and Prendinger, H. 2019. UAV-based situational awareness system using deep learning. *IEEE access*, 7: 122583–122594.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Herzig, R.; Ben-Avraham, E.; Mangalam, K.; Bar, A.; Chechik, G.; Rohrbach, A.; Darrell, T.; and Globerson, A. 2022. Object-region video transformers. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3148–3159.

Joao, C.; and Andrew, Z. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 556–560.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. In *Advances in neural information processing systems*, 18661–18673.

Kondratyuk, D.; Yuan, L.; Li, Y.; Zhang, L.; Tan, M.; Brown, M.; and Gong, B. 2021. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of IEEE/CVF conference on computer vision and pattern recognition*, 16020–16030.

Koohzadi, M.; and Charkari, N. M. 2017. Survey on deep learning methods in human action recognition. *IET Computer Vision*, 11(8): 623–632.

Kothandaraman, D.; Guan, T.; Wang, X.; Hu, S.; Lin, M.; and Manocha, D. 2022. FAR: Fourier Aerial Video Recognition. In *European Conference Computer Vision*, 657–676.

Kothandaraman, D.; Lin, M.; and Manocha, D. 2022. Differentiable frequency-based disentanglement for aerial video action recognition. *arXiv preprint arXiv:2209.09194*.

Li, Q.; Qiu, Z.; Yao, T.; Mei, T.; Rui, Y.; and Luo, J. 2016. Action recognition by learning deep multi-granular spatio-temporal video representation. In *ACM on International Conference on Multimedia Retrieval*, 159–166.

Li, T.; Liu, J.; Zhang, W.; Ni, Y.; Wang, W.; and Li, Z. 2021. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16266–16275.

Lin, J.; Gan, C.; and Han, S. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7083–7093.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3202–3211.

Patrick, M.; Campbell, D.; Asano, Y.; Misra, I.; Metze, F.; Feichtenhofer, C.; Vedaldi, A.; and Henriques, J. F. 2021. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in Neural Information Processing Systems*, 12493–12506.

Perera, A. G.; Law, Y. W.; and Chahl, J. 2019. Drone-action: An outdoor recorded drone video dataset for action recognition. *Drones*, 3(4): 82.

Reddy, A. V.; Shah, K.; Paul, W.; Mocharla, R.; Hoffman, J.; Katyal, K. D.; Dinesh, V.; de Melo, C. M.; and Chellappa, R. 2022. RoCoG-v2. Under Review.

Shi, G.; Fu, X.; Cao, C.; and Zha, Z.-J. 2023. Alleviating Spatial Misalignment and Motion Interference for UAV-based Video Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 193–202.

Shu, X.; Xu, B.; Zhang, L.; and Tang, J. 2022. Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition. *Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7559–7576.

Shu, X.; Zhang, L.; Qi, G.-J.; Liu, W.; and Tang, J. 2021. Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3300–3315.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *proceedings of the IEEE international conference on computer vision*, 4489–4497.

Wang, H.; and Schmid, C. 2013. Action recognition with improved trajectories. In *proceedings of the IEEE International Conference on Computer Vision*, 3551–3558.

Wang, J.; Yang, X.; Li, H.; Liu, L.; Wu, Z.; and Jiang, Y.-G. 2022a. Efficient video transformers with spatial-temporal token selection. In *European Conference on Computer Vision*, 69–86.

Wang, P.; Wang, X.; Wang, F.; Lin, M.; Chang, S.; Li, H.; and Jin, R. 2022b. Kvt: k-nn attention for boosting vision transformers. In *European Conference on Computer Vision*, 285–302.

Wang, X.; Xian, R.; Guan, T.; de Melo, C. M.; Nogar, S. M.; Bera, A.; and Manocha, D. 2023. Aztr: Aerial video action recognition with auto zoom and temporal reasoning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 1312–1318. IEEE.

Xian, R.; Wang, X.; Kothandaraman, D.; and Manocha, D. 2024. Pmi sampler: Patch similarity guided frame selection for aerial action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6982–6991.

Xian, R.; Wang, X.; and Manocha, D. 2024. Mitfas: Mutual information based temporal feature alignment and sampling for aerial video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6625–6634.

Zhao, G.; Lin, J.; Zhang, Z.; Ren, X.; Su, Q.; and Sun, X. 2019. Explicit sparse transformer: Concentrated attention through explicit selection. *arXiv preprint arXiv:1912.11637*.

Zhi, Y.; Tong, Z.; Wang, L.; and Wu, G. 2021. Mgsampler: An explainable sampling strategy for video action recognition. In *Proceedings of the IEEE/CVF International conference on computer Vision*, 1513–1522.