PMI Sampler: Patch Similarity Guided Frame Selection For Aerial Action Recognition

Ruiqi Xian, Xijun Wang, Divya Kothandaraman, Dinesh Manocha
University of Maryland - College Park
{rxian, xijun, dkr, dmanocha}@umd.edu

Abstract

We present a new algorithm for the selection of informative frames in video action recognition. Our approach is designed for aerial videos captured using a moving camera where human actors occupy a small spatial resolution of video frames. Our algorithm utilizes the motion bias within aerial videos, which enables the selection of motion-salient frames. We introduce the concept of patch mutual information (PMI) score to quantify the motion bias between adjacent frames, by measuring the similarity of patches. We use this score to assess the amount of discriminative motion information contained in one frame relative to another. We present an adaptive frame selection strategy using shifted leaky ReLu and cumulative distribution function, which ensures that the sampled frames comprehensively cover all the essential segments with high motion salience. Our approach can be integrated with any action recognition model to enhance its accuracy. In practice, our method achieves a relative improvement of 2.2 - 13.8% in top-1 accuracy on UAV-Human, 6.8% on NEC Drone, and 9.0% on Diving48 datasets. The code is available at https://github.com/Ricky-Xian/PMI-Sampler.

1. Introduction

The applications of unmanned aerial vehicles (UAVs) have been expanding rapidly into search and rescue, agriculture, security, surveillance, etc. This is giving rise to many challenging problems including detection, re-identification, tracking, and recognition. Among these tasks, aerial video action recognition is regarded as one of the most difficult problems. Despite the success of deep learning-based methods in video action recognition on ground camera footage [10, 20, 30], current methods do not result in high accuracy on aerial videos.

Aerial video data are often captured by a camera mounted on a moving UAV in oblique or overhead angles. The resulting footage features human actors that appear significantly smaller (typically less than 10% pixels), due to the high camera altitude, with a large portion of the video frame dominated by the background information. The size and scale of the human actor may vary considerably, owing to changes in flying altitude during video data collection. Furthermore, the continuous movement of the UAV can cause the camera viewing angle shifting, resulting in motion blur and occlusion. These factors collectively make the development of accurate algorithms for aerial action recognition a challenging task.

The decrease in performance of generic video recognition models on aerial data can be reduced through the use of motion-guided frame sampling during training [23, 46]. However, current deep learning-based action recognition methods mostly use fixed hand-crafted sampling techniques for video analysis [38, 41, 43]. Typically, frames are randomly sampled in a uniform manner or successively with a fixed stride from the original video. This fixed sampling strategy can be sub-optimal for several reasons. First, the motion duration varies for different videos and actions, and fixed sampling may not capture the entire motion duration, potentially overlooking some useful information. Second, the sampling approach should prioritize discriminative frames over redundant or uninformative background frames, as not all frames are equally useful in terms of recognition.

Recently, some techniques [13, 15, 21, 26, 45] have been proposed for frame selection by modeling it as a decision making task. Typically, these methods employ a learning-based module to sequentially select more informative frames or to conditionally exit early. While these methods have shown promising results, their performance heavily relies on the training data and may not easily transfer to unseen domains. Unfortunately, the scarcity of labeled aerial videos, coupled with the challenges of collection and annotation, makes the task of training such modules more difficult. It turns out that the number and size of UAV video datasets is far fewer and smaller than those available for ground video datasets. For instance, the Kinetics dataset contains 650k videos, while the UAV-Human dataset has only 20k videos. Additionally, these learning-based methods are primarily de-

Figure 1. Sample eight frames from typical videos in Diving48 and UAV-Human. Compared with the state-of-the-art, MG Sampler [48], our approach can better represent the motion distribution and provides an easier way to distinguish motion salient frames in videos. PMI Sampler is more robust to noises and outliers and can better handle the background changes caused by the moving camera.

signed for untrimmed videos, and adapting them to trimmed videos poses additional challenges. Another alternative [49] is to use a statistical model to represent the motion bias between frames, and devise an adaptive sampling strategy for frame selection based on motion information distribution along the temporal domain. However, this formulation fails to account for the distinct features of aerial video data, such as small resolution, multi-scale, and moving camera.

Main Contribution: We present a novel frame selection scheme for aerial action recognition. Our approach is general and is designed to address some of the challenges in aerial data. We utilize the similarity between patches to assess the amount of discriminative motion information contained in the aerial videos, and ensure that more informative frames can be selected for video representation. Our method can be combined with any recognition model to obtain improved accuracy in terms of aerial action recognition. The novel components of our work includes:

1. We introduce patch mutual information (PMI) score to leverage the motion bias in the aerial videos. PMI score quantifies how much discriminative motion information is contained in one frame given another by measuring the similarity of frame patches via mutual information calculation.

2. We propose an adaptive frame selection strategy based on shifted Leaky ReLu and cumulative distribution function. Our formulation enhances the motion bias between adjacent frames, making it easier to distinguish the motion-salient frames in videos. Furthermore, our method is designed in a direct plug-in manner to avoid complex training.

We evaluate the effectiveness of our method in three aerial datasets and experimental results show that our approach consistently outperforms current state-of-the-art by large margins in terms of top-1 accuracy. Practically, we demonstrate a relative improvement of 2.2 - 13.8% on UAV-Human [24], 6.8% on NEC Drone [4], and 9.0% on Diving48 datasets [25].

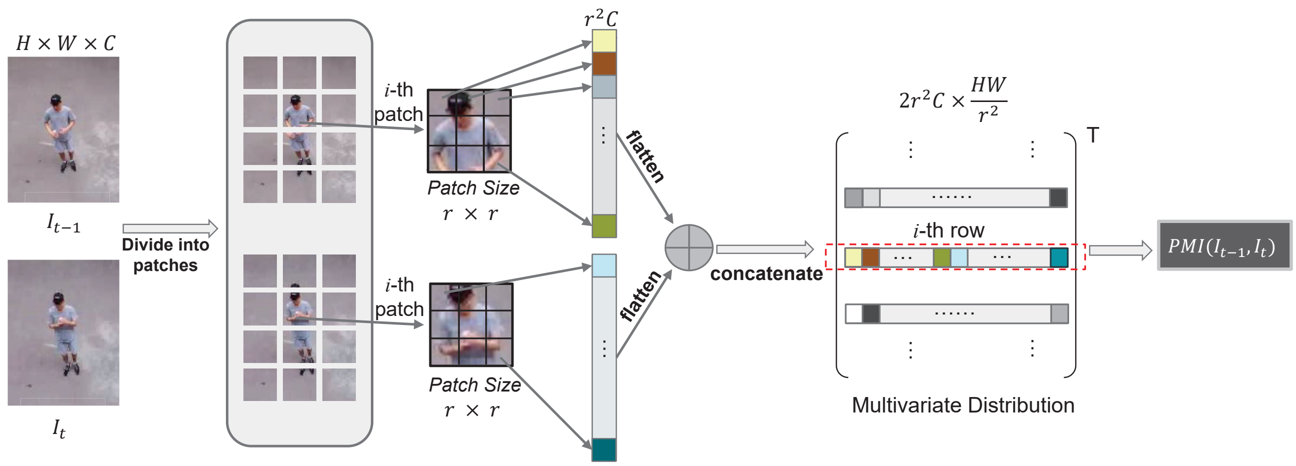

Figure 2. **Patch Mutual Information formulation:** Given a pair of adjacent frames I_{t-1} and I_t in size $H \times W \times C$, we first divide them into patches with size $r \times r \times C$ and map each patch to a d-dimensional vector, where $d = r^2C$. We concatenate the two d-dimensional vectors representing the corresponding patches from two frames, the resulted vector is in $2d$ space. We perform the same operations for each patches and further concatenate these $2d$ vectors to get the multivariate distribution of size $2d \times \frac{HW}{r^2}$ for this image pair. Then, we calculate the covariance matrix of the multivariate distribution, compute the approximated entropies using Eq 5 and eventually get the patch mutual information following the Eq 3

2. Related Work

2.1. Action Recognition for Aerial Videos

The accuracy of action recognition on ground-camera video datasets has increased due to recent advancements in deep learning techniques. However, the current method cannot demonstrate a similar level of accuracy on videos captured using UAV cameras [32]. For aerial videos, [14], [29], [28], [31], [1], [12], [31] utilize 2D convolutional neural networks (such as ResNet and MobileNet) as the foundational models for single-frame classification, and subsequently merge the outcomes of all video frames. Other methods [2], [34], [35] utilize two-stream convolutional neural networks (CNNs) to leverage both human motion and appearance attributes for improved action recognition. [4], [6], [24], [31], [39] employ the I3D network [3] to capture the spatial-temporal features of both human agents and their surroundings. [44] proposes a framework leveraging CNNs and attention mechanisms for aerial action recognition on both edge devices and decent GPUs. [22] introduced an attention mechanism based on the Fourier transform to emphasize motion salience. Our proposed PMI Sampler is complimentary and could incorporate with aforementioned method to improve the overall recognition accuracy.

2.2. Frame Sampling

For some deep learning-based methods [3, 9, 40], the frame sequence used for training is obtained by randomly picking a fixed number of consecutive frames in the video. Other methods [10, 43] use a uniform sampling strategy, where frames are evenly sampled along the video's temporal domain. These two sampling techniques are commonly used for action recognition models. However, they do not exploit the motion bias between frames and do not consider the video characteristics corresponding to different human actions. There are also some learning-based frame selection methods [8, 13, 21, 45, 47]. FastForward [8] employs reinforcement learning for planning frame skipping and early stop decisions. Adaframe [45] utilizes a policy gradient-trained LSTM, enhanced with a global memory for frame selection. SCSampler [21] and Listen to Look [13] utilize audio as an additional modality to exploit the natural semantic correlation between audios and frames. However, those methods mainly focus on untrimmed videos. They require large amounts of training samples and involve complex training procedures.

MGSampler [49] leverages the temporal variations and use RGB difference between two adjacent frames to estimate the motion salience for each frame. However, such an approach may not be robust. Considering the challenges of aerial data (e.g. small resolution for human actors), their formulation may not be accurate due to the large number of outliers and noises belonging to the background information. In order to address these issues, we present a motion information representation technique using patch mutual information between frames and introduce a new sampling strategy based on the PMI score and shifted leaky ReLu.

2.3. Similarity Measure between Images

Numerous similarity metrics have been suggested for image analysis, with Euclidean distance being a common choice [48]. Nevertheless, for UAV videos, Eu-

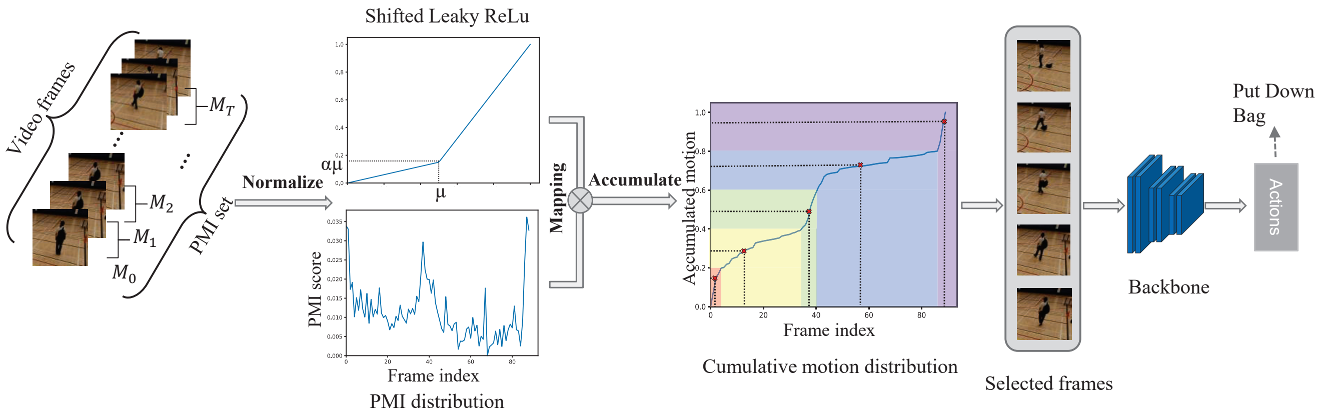

Figure 3. **PMI Sampler:** Given a sequence of video frames, we compute the patch mutual information score for each pair of adjacent frames. We further remap the PMI scores using shifted Leaky ReLu to enhance the motion distribution. Finally, we accumulated the PMI scores using cumulative distribution function and segment the video into N parts where N is the number of samplings required for training. We randomly select one frame from each segments, constitute them into a sequence and feed them to the recognition backbones for action classification.

clidean distance struggles due to pronounced background noise and frame instability. Cosine similarity is used for high-dimensional data but overlooks pixel value magnitude [16]. Mutual information is a well-adopted similarity measure [27, 42], especially in medical registration [19, 36]. It boasts resistance to outliers, yielding smooth cost functions for optimization [5]. However, it lacks geometry consideration, solely focusing on pixel values, and neglecting spatial pixel relationships. NMI, which is a variation of Mutual Information, overlooks the spatial correlation between pixels and incurs higher computational costs. PSNR, derived from Mean Square Error (MSE), primarily focuses on pixel-level comparisons, is sensitive to dominant background changes encountered in aerial videos. Similarly, SSIM, a widely employed similarity measure, assesses the luminance, contrast, and structure of images but proves to be highly sensitive to structural variations such as rotations and shifts, which are frequently observed in aerial videos. Russakoff et al. [37] propose RMI, integrating spatial data with mutual information, but it's computationally expensive. Inspired by these, we introduce patch mutual information, an RMI extension, simpler to implement yet provides enhanced accuracy.

3. Our Approach: PMI Sampler

In this section, we present the details of our proposed patch similarity guided frame selection strategy. Specifically, we first introduce the concept of mutual information in Section 3.1. Next, we present the key component of our approach, including patch mutual information (PMI) score in Section 3.2. Finally, we present the overall pipeline of our proposed PMI Sampler based on shifted leaky ReLu and cumulative distribution function in Section 3.3.

3.1. Mutual Information

Mutual information (MI) is a concept in information theory that essentially measures the amount of information given by one variable when observing another variable. Mutual information is highly correlated with entropy and joint entropy. The entropy is a measure of the uncertainty of a random variable and the joint entropy examines the overall complexity of all possible outcomes given both random variables. Specifically, given two discrete random variables X and Y with alphabet $\mathcal{X}$ and $\mathcal{Y}$. Their probability mass functions (PMFs) are denoted as $p_X(x)$ and $p_Y(y)$, the entropy of X,$H(X)$, and the joint entropy of X and Y, $H(X, Y)$, can be calculated as:

$$H(X) = -\sum_{x\in\mathcal{X}} p_X(x)\log p_X(x). \tag{1}$$

$$H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{XY}(x,y)\log p_{XY}(x,y). \tag{2}$$

The mutual information between X and Y is defined as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \tag{3}$$

If we plug Eq 1 and Eq 2 into Eq 3, we can get the calculation of mutual information as follows:

$$I(X;Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{XY}(x,y)\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}. \tag{4}$$

The equation 4 further suggests that mutual information essentially measures the distance between the real joint distribution $p_{XY}(x, y)$ and the distribution under the assumption of complete independence of $p_X(x)p_Y(y)$, which makes it a very nature measure of dependence [18].

Mutual information can be used to measure the similarity between images [27, 42]. The probability distributions (in Eq 4) associated with images are normally approximated using respective marginal and joint histograms. One set of pixel co-occurrence in the joint distribution is represented by one entry in a two-dimensional joint histogram.

Regional mutual information [37] is an extension of mutual information. The intuition of RMI is to embed the spatial information of the images into mutual information calculations. Instead of standard pixel-wise mutual information, RMI represents a pixel as a multi-dimensional vector that consists of not only the pixel itself but also its neighboring pixels. However, the way that representing pixels with their neighborhood essentially makes extra focus on the local details. It may be beneficial for generic camera datasets, but considering that most of the pixels belong to the background in aerial data and the background details are redundant for action recognition, RMI is not a good choice to measure the motion information in aerial videos.

3.2. Patch Mutual Information (PMI) score

Inspired by regional mutual information (RMI), we propose patch mutual information (PMI). Instead of mapping pixels into multi-dimensional space, we divide the image into patches and map each patch into a multi-dimensional points and then calculate the mutual information between corresponding patches in adjacent frames. In this manner, we can encode the spatial relationships of patches in the mutual information calculation without giving too much focus on the redundant background noises, which yields to better motion information representation for aerial videos.

Entropy approximation We can exactly calculate the patch mutual information in the traditional manner that mentioned in Section 3.1, but now with d-dimensional histograms for marginal probability distribution and $2d$-dimensional histograms for joint distribution. However, such operation seems reasonable but is hard to implement in practice. In fact, we can make the calculation easier by assuming that the multivariate distribution of the image is normally distributed [37] and it is well supported by the m-dependence variable concept proposed in [17].

For a normally distributed set of points $P = [p_1, p_2, \cdots, p_N]^T$, $p_1, p_2, \cdots, p_N \in R^d$, the entropy can be calculated as:

$$H(P) = \frac{1}{2} log((2\pi e)^d det(var(P))). \quad (5)$$

where $var(P)$ is the covariance matrix of P. Let $E(P)$ be the expected value of P, i.e, the mean of P, covariance matrix of P is defined as:

$$var(P) = E[(P - E[P])(P - E[P])^T], \quad (6)$$

Patch mutual information formulation We here describe the detailed procedures to get the multivariate distribution for a pair of images and calculate the patch mutual information between these two images.

As shown in Figure 2, given frame I_{t-1} and I_t, we divide each frame into patches with square length r. For each patch pair, we flatten the patches into vectors in d-dimensional space where $d = Cr^2$ and concatenate them to a vector in R^{2d}. For each pair of frames in $R^{H\times W\times C}$, we will have $N = \frac{HW}{r^2}$ patches (we simply ignore the pixels along the edges as they will not have a significant effect on the final entropy). Since each pair of patches has been mapped into a $2d$ vector, we then have a multivariate distribution of those N points by further concatenating the $2d$ vectors into a $2d \times N$ matrix, $P_t = [p_1, p_2, \cdots, p_N]$.

We calculate the covariance matrix of P_t by first centering all the N points to their means and then compute the covariance of P_t following Eq. 6. The covariance matrix of P_t is denoted as Ω:

$$\Omega = \frac{1}{N}(P_t - \frac{1}{N}\sum^{N} p_i)(P_t - \frac{1}{N}\sum^{N} p_i)^T. \quad (7)$$

We estimate the entropy of these points $H(P_t)$ using Eq 5. Since we concatenate the patches from two frames to form $P_t \in R^{2d\times N}$, $H(P_t)$ is essentially the approximation of the joint entropy between frame I_{t-1} and I_t. However, we can easily obtain the covariance matrices of each frame. Given that P_t is a combination of two d-dimensional vectors, the computation of the covariance matrix of P_t also results in the covariance matrices of these 2 subsets of points. In practice, the covariance of frame I_{t-1}, denoted as Ω_{t-1}, corresponds to the $d \times d$ matrix in the top left of Ω and Ω_t is the $d \times d$ matrix in the bottom right representing the covariance of frame I_t. Moreover, the marginal entropies, $H(\Omega_{t-1})$ and $H(\Omega_t)$, can be computed using the same formulation in Eq 5.

Finally, we obtain the patch mutual information in a pair of frames I_{t-1} and I_t, denoting as M_t, following the Eq 3:

$$M_t = PMI(I_{t-1}, I_t) = H(\Omega_{t-1}) + H(\Omega_t) - H(\Omega). \quad (8)$$

Patch mutual information score Note that, PMI holds the properties of standard mutual information and is always greater than or equal to 0. In case, the two frames are complete independent from each other, then $PMI = 0$. However, since video contents are always consistent and have some coherence between the frames, the two successive frames are always correlated and dependent. Therefore, PMI between adjacent frames will always be greater than zero, $M_t > 0$ for $t \in 1, 2, \cdots, T$. For frame at time 0, we simply define $M_0 = 0$.

When the two frames are more discriminative, the dependence relationship weakens and PMI gets smaller, indicating there is potentially more motion information contained in the current frame. The scale and range of the PMI in different videos also varies. To better analyze the motion information distribution, we inversely remap the PMI M_t and normalize it with the l_1 norm:

$$M'_t = \max_{i \in T}(M_i) - M_t, \tag{9}$$

$$\hat{M}_t = \frac{M'_t}{\sum_T M'_t}. \tag{10}$$

Eventually, we obtain the PMI score $\hat{M}_t$. PMI score represents the discriminative motion information contained in current frame given the previous frame. The higher PMI score indicates there existing salient motion in current frame.

3.3. PMI Sampler: Frame Selection

We present a novel frame selection strategy based on shifted leaky ReLu and cumulative distribution function (CDF). The intuition behind our method is to sample a sequence of frames that contains as much discriminative motion information as possible. Similar to MG Sampler [48], our method also utilize a temporal segmentation scheme and adaptively selects frames according to the motion information distribution over the entire video.

As shown in Figure 3, given a video contain T frames, we calculate PMI between adjacent frames and generate a set of PMI scores, $S = \{\hat{M}_0, \hat{M}_1, \cdots, \hat{M}_T\}$ with mean μ. For each element $\hat{M}_t \in [0, 1], t \in \{0, 1, \cdots, T\}$, we map it based on the following equation:

$$M^*_t = \begin{cases} \alpha \hat{M}_t & \hat{M}_t \leq \mu, \\ \frac{1-\alpha\mu}{1-\mu}(\hat{M}_t - 1) + 1 & \hat{M}_t > \mu. \end{cases} \tag{11}$$

where α is a hyper parameter to control the smoothness. This function is very much similar to Leaky ReLu, but with the origin shifts to $(\mu, \alpha\mu)$ and the function is constrained by the origin and $(1, 1)$. Therefore, we call it as *shifted Leaky ReLu*. As shown in Figure 4, it polarizes the PMI scores and makes the motion-salient frames more distinct. We further examine the impact of α in Section 4.3.

Then, we normalize M^*_t again using l_1 norm and accumulate the the remapped PMI score to get the cumulative motion distribution of the video:

$$F_T(t) = \sum_{i \leq t} M^*_i. \tag{12}$$

The constructed cumulative distribution of motion information is shown in Fig 3, where the X-axis represents the frame index and Y-axis stands for the motion information accumulation up to the current frame.

Finally, we divide the video frames into N parts with interval $1/N$ on the Y-axis. The frames with index in the corresponding interval on the X-axis will be clustered as one segment. Considering that the X-axis value may not be an integer, we choose the closest integer value instead.

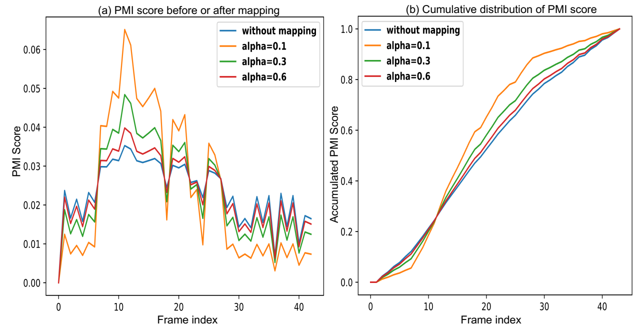

Figure 4. The PMI score and cumulative distribution before or after the shifted Leaky ReLu mapping under different value of α. After mapping, motion salient frames become more distinct.

Then, we will randomly select one frame from each segment and constitute a frame sequence to represent the video for the recognition. As shown in Fig 3 and Fig 1, our approach can select more frames during periods with higher motion salience and less frames in motion static segments, enabling more discriminative motion information to be conveyed to the recognition model.

4. Results

In this section, we present the experimental results of our approach along with the state-of-the-art methods on three aerial datasets: UAV-Human [24], NEC Drone [4] and Diving48 [25]. We also conduct ablation studies on the effects of patch mutual information, shifted leaky ReLu, and patch size. Our approach can incorporate with any existing recognition backbones, we use X3D [10] as our default recognition backbone model unless otherwise specified. The implementation details are included in supplementary material.

4.1. Results on UAV-Human

UAV Human [24] is regarded as the largest and most comprehensive dataset of UAV-based human behavior understanding data to date. The collection includes 22,476 high-definition videos captured in various indoor and outdoor settings, encompassing a broad range of lighting and weather conditions. The videos showcase dynamic backgrounds and feature diverse UAV motions and flying 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.

We evaluate the performance of our proposed PMI Sampler on UAV-Human along with current state-of-the-art methods. The results are shown in Table 1. All the videos are pre-processed with the same procedures in MITFAS [46], and followed by data augmentation that consistent with X3D [10]. Our method demonstrates a relative improvement in top-1 accuracy over the current state-of-the-art methods by $2.2 - 13.8\%$, with regards to different settings in number

Method	Frames Number	Input Size	Init.	Top-1 (%) Acc. ↑
FAR [22]	8	540 × 540	None	28.8
MITFAS [46]	8	540 × 540	None	38.4
Ours	8	540 × 540	None	**39.7**
X3D-M [10]	8	540 × 540	Kinetics	36.6
FAR [22]	8	540 × 540	Kinetics	38.6
DiffFAR [23]	8	540 × 540	Kinetics	41.9
Ours	8	540 × 540	Kinetics	**47.7**
FAR [22]	8	620 × 620	Kinetics	39.1
MITFAS [46]	8	620 × 620	Kinetics	46.6
Ours	8	620 × 620	Kinetics	**52.0**
X3D-M [10]	16	224 × 224	Kinetics	30.6
FAR [22]	16	224 × 224	Kinetics	31.9
AZTR [44]	16	224 × 224	Kinetics	47.4
MITFAS [46]	16	224 × 224	Kinetics	50.8
MG Sampler [48]	16	224 × 224	Kinetics	53.8
Ours	16	224 × 224	Kinetics	**55.0**

Table 1. **Results on UAV-Human.** We demonstrate relative improvements in the top-1 accuracy by $2.2 - 13.8\%$ over previous state-of-the-art methods. Our method outperforms the state-of-the-art methods under different settings, which further indicates the benefits of our proposed PMI Sampler.

Method	Frames	Backbone	Diving48 Top-1 (%)	NEC Drone Top-1 (%)
Random [10]	16	X3D-M	71.1	52.0
Uniform [10]	16	X3D-M	73.5	55.4
MG Sampler [48]	16	X3D-M	74.6	58.5
K-centered [33]	16	ViT [7]	72.5	36.3
Ours	16	X3D-M	**81.3**	**62.5**

Table 2. **Results on Diving48 and NEC-Drone.** Our method relatively improves the top-1 accuracy by 9.0% on Diving48, by 6.8% on NEC Drone.

of frames, input frame size, and model initialization.

4.2. Results on Diving48 and NEC-Drone

NEC Drone [4] is an indoor video dataset that features 5,250 videos depicting 16 distinct actions performed by 19 actors. The videos were captured using a UAV flying at low altitude over a basketball court. In contrast to the UAV Human dataset, the lighting conditions in NEC Drone are more consistent; however, the dataset is plagued by noise due to light reflections. Diving48 [25] is a comprehensive video dataset that offers a fine-grained analysis of competitive diving, free from any significant biases towards static or short-term motion representations. It comprises approximately 18,000 trimmed video clips, each depicting one of 48 unambiguous dive sequences. Although it is not a UAV-captured dataset, the majority of the videos are captured by cameras at high altitudes and oblique angles with dynamic movement.

We compare our method with other state-of-the-art methods on Diving48 and NEC-Drone. The frames are extracted

Similarity measure	UAV-Human subset Top-1 Acc (%)	Time cost per frame (ms)
Euclidean Distance	56.4	1.7
Cosine Similarity	55.0	2.5
Mutual Information	58.4	10.6
Regional Mutual Information (RMI)	59.2	20.2
Normalized Mutual Information (NMI)	58.8	32.8
Peak Signal-to-Noise Ratio (PSNR)	57.0	3.7
Structural Similarity Index Measure (SSIM)	57.7	79.9
Patch Mutual Information	**59.8**	**4.5**

Table 3. **Comparison between different similarity measures**. Our proposed Patch Mutual Information(PMI) outperforms other measures on UAV-Human.

Mapping function	NEC Drone Top-1 Acc (%)	Diving48 Subset Top-1 Acc (%)
without Mapping	59.3	58.8
Quadratic	60.3	63.4
Sigmoid	61.4	65.7
Softmax	58.4	53.9
Tanh	60.9	56.6
ReLu	60.5	60.6
Shifted Leaky ReLu	**62.5**	**66.3**

Table 4. **Comparison between different mapping functions.** Shifted Leaky ReLu can better map the motion information distribution and make it easier to distinguish frames containing more motion information.

from raw videos and augmented as in X3D [10]. The baseline methods are uniform and random samplings. As shown in Table 2, on diving48, our method achieves 10.6 - 14.3% relative improvement in top-1 accuracy over the baseline methods and relative 9.0% over the SOTA. On NEC Drone, PMI Sampler outperforms the baseline methods by 12.8 - 20.2% and improves by 6.8% over the SOTA, relatively.

4.3. Ablation Studies

We conduct ablation studies on the effectiveness of patch mutual information and shifted leaky ReLu. We also explore the impact of the hyper-parameter α in the shifted leaky ReLu functions as well as the size of the patches. We generate subsets of UAV-Human and Diving48 by randomly choosing 30% videos per class from the original datasets and denote them as UAV-Human subset ($\sim$ 6k videos) and Diving48 subset ($\sim$ 5.8k videos).

Effectiveness of patch mutual information: As mentioned in Section 2.3, there are many other similarity measurements like Euclidean distance, cosine similarity, . We conduct experiments to test the results generated based on different similarity measures. As shown in table 3, our proposed PMI outperforms other similarity measures.

Effectiveness of shifted Leaky ReLu: After getting the PMI score for the video, we remap it using shifted Leaky ReLu. It polarizes the motion information distribution and makes it easier to distinguish the motion-salient frames in aerial videos, see Figure 4. However, there are many other

Method	Frames	Input Resolution	Training Time per Epoch (s) ↓	Training Time per Video (ms) ↓	Diving48 Top-1(%) ↑	Diving48 Top-5(%) ↑
Random	16	224×224	196.6	13.1	71.1	94.8
Uniform	16	224×224	191.4	12.7	73.5	95.1
MG Sampler	16	224×224	287.4	19.2	74.6	95.0
PMI Sampler **(Ours)**	**8**	224×224	248.3 **(-39.1)**	16.5 **(-2.7)**	75.0 **(+0.4)**	95.6 **(+0.6)**
PMI Sampler **(Ours)**	**12**	224×224	276.8 **(-10.6)**	18.4 **(-0.8)**	77.7 **(+3.1)**	97.5 **(+2.5)**
PMI Sampler **(Ours)**	16	$\mathbf{172 \times 172}$	294.8 **(+7.4)**	19.6 **(+0.4)**	79.0 **(+4.4)**	97.7 **(+2.7)**
PMI Sampler **(Ours)**	16	224×224	295.1 **(+7.7)**	19.6 **(+0.4)**	81.3 **(+6.7)**	97.7 **(+2.7)**

Table 5. **Training efficiency.** PMI Sampler achieves higher accuracy with less spatial/temporal information, leading to better speed-accuracy tradeoffs. We show results for fewer frames (8 and 12) and smaller input sizes (172×172).

	α	0	0.1	0.2	**0.3**	0.4	0.5	0.6	0.7
Diving48 subset	Top-1 (%)	60.6	65.4	64.7	**66.3**	66.2	65.4	65.6	65.4
	Top-5 (%)	92.6	92.8	94.4	**94.9**	94.6	93.3	94.5	94.6
UAV-Human subset	Top-1 (%) ↑	53.7	58.1	59.8	**59.8**	59.7	58.7	59.5	59.3
	Top-5 (%) ↑	83.8	85.3	**86.5**	86.3	85.3	86.2	85.5	86.0
NEC-Drone	Top-1 (%) ↑	54.1	57.9	58.0	58.2	60.7	61.2	**62.5**	61.7
	Top-5 (%) ↑	85.4	89.5	88.6	89.3	89.5	**90.0**	89.6	89.5

Table 6. **Impact of the hyper-parameter** α**.** The dynamic camera movement in UAV-Human and Diving48 videos necessitates a lower $\alpha(\sim 0.3)$ to differentiate real motion from background noise. In contrast, NEC-Drone videos, captured by a hovering UAV, require a higher $\alpha(\sim 0.6)$ to preserve the motion info distribution.

mapping functions we can use, like sigmoid, softmax, tanh functions. We conduct ablation experiments on different mapping functions, and the results are shown in Table 4, which demonstrate that shifted Leaky ReLu is the better choice for aerial videos.

Impact of the hyper-parameter in shifted leaky ReLu: We also conduct experiments to explore the impact of the hyper-parameter α in the shifted leaky ReLu function. Results are shown in Table 6. When α is smaller, the distance between small PMI scores and large PMI scores will be enlarged, yielding to steeper cumulative motion information distribution, see Figure 4. With α larger, such distance will be reduced and the cumulative distribution will be smoother.

Impact of the patch size: We compare the results from PMI scores that were generated using different sizes of patches. As we mentioned in Section 3.2, the entropy approximation is based on the assumption that the multivariate distribution of the image is normally distributed, and it is well supported by the m-dependence variable concept. Such approximation is closer to the real entropy when the patch size is larger. However, if the patch size is larger, the time for the calculation will be more expensive. As shown in Table 7, we find out that when patch size is 7×7, we achieve the best trade-off between accuracy and time cost. Therefore, we divide the images into 7×7 patches for all the other experiments in this paper unless further specified.

Training efficiency: The training time on the Diving48 dataset is presented in Table 5. Similar to the MG Sampler, our approach computes the PMI in the pre-processing phase to avoid redundant computations for the same set of frames during training. PMI Sampler achieves improved accuracy without significantly increasing the training time. Also, compared to MG Sampler, our proposed PMI Sampler achieves better accuracy with less spatial or temporal information (fewer frames and smaller input size).

Size of Patch ($r \times r$)	UAV-Human Top-1 Accuracy (%)	Time cost per frame (ms)
3×3	53.8	5.4
5×5	54.6	6.8
7×7	**55.0**	**9.2**
11×11	52.3	17.9
21×21	51.9	67.8

Table 7. **Impact of the patch size.** A better tradeoff between accuracy and time cost is achieved with patch size 7×7.

Dataset	**SthSthV2**		**UCF101**		**HMDB51**	
	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)
Uniform	57.5	84.3	94.3	99.3	64.6	88.1
MG Sampler	59.2	85.2	94.6	99.2	64.9	87.9
Ours (Patch 7×7)	59.7 **(+0.5)**	85.2 **(+0)**	94.4 **(-0.2)**	99.6 **(+0.4)**	64.5 **(-0.4)**	87.2 **(-0.7)**
Ours (Patch 3×3)	60.3 **(+1.1)**	85.8 **(+0.6)**	95.1 **(+0.5)**	99.5 **(+0.3)**	65.5 **(+0.6)**	88.1 **(+0.2)**

Table 8. **Results on SthSthV2, UCF101, HMDB51**. Given that the majority of videos in these datasets feature fixed cameras capturing close-range scenes with prominent human actors and minimal background changes, utilizing smaller patches (3×3) can effectively capture finer details of the actions, leading to improved accuracy.

Results on ground datasets: We evaluate our proposed PMI Sampler on general ground camera datasets like SomethingSomething V2, UCF101 and HMDB51. UCF101 and HMDB51 both have 3 train-test splits, we report the average results here with X3D as the backbone. As shown in Table 8, PMI Sampler also outperforms the current state-of-the-art on these datasets.

More ablation studies are shown in Appendix. B and C. We also include a detailed analysis in Appendix. D and more visualization results in Appendix. E.

5. Conclusion, Limitations and Future Work

In this paper, we present a novel frame selection method for aerial video action recognition. We first introduce patch mutual information (PMI) score to represent the motion information between adjacent frames by measuring the similarity of frame patches via mutual information calculation. Then, we propose an adaptive frame selection strategy based on shifted Leaky ReLu and cumulative distribution function, which ensures that the sampled frames comprehensively cover all the essential segments with high motion salience. Our evaluations on multiple datasets illustrate the effectiveness of our method. Even though we improved the SOTA methods, the absolute accuracy on UAV-Human is not very high, and we will further explore the backbone architecture design to improve the accuracy.

Acknowledgement This work was supported in part by ARO Grants W911NF2110026, W911NF2310046, W911NF2310352 and Army Cooperative Agreement W911NF2120076

References

[1] Oscar L Barbed, Pablo Azagra, Lucas Teixeira, Margarita Chli, Javier Civera, and Ana C Murillo. Fine-grained pointing recognition for natural drone guidance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1040–1041, 2020. 3

[2] Mohammadamin Barekatain, Miquel Martí, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. Okutama-action: An aerial view video dataset for concurrent human action detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 28–35, 2017. 3

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017. 3, 11

[4] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1717–1726, 2020. 2, 3, 6, 7, 11, 14, 15

[5] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. 4

[6] Ugur Demir, Yogesh S Rawat, and Mubarak Shah. Tinyvirat: Low-resolution video action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7387–7394. IEEE, 2021. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[8] Hehe Fan, Zhongwen Xu, Linchao Zhu, Chenggang Yan, Jianjun Ge, and Yi Yang. Watching a small portion could be as good as watching all: Towards efficient video classification. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 705–711. International Joint Conferences on Artificial Intelligence Organization, 7 2018. 3

[9] Hehe Fan, Zhongwen Xu, Linchao Zhu, Chenggang Clarence Yan, Jianjun Ge, and Yi Yang. Watching a small portion could be as good as watching all: Towards efficient video classification. In *International Joint Conference on Artificial Intelligence*, 2018. 3

[10] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 203–213, 2020. 1, 3, 6, 7, 11

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 11

[12] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Predicting the future: A jointly learnt model for action anticipation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5562–5571, 2019. 3

[13] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020. 1, 3

[14] Ruben Geraldes, Artur Goncalves, Tin Lai, Mathias Villerabel, Wenlong Deng, Ana Salta, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. Uav-based situational awareness system using deep learning. *IEEE Access*, 7:122583–122594, 2019. 3

[15] Amir Ghodrati, Babak Ehteshami Bejnordi, and Amirhossein Habibian. Frameexit: Conditional early exiting for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15608–15618, 2021. 1

[16] Jiun Tian Hoe, Kam Woh Ng, Tianyu Zhang, Chee Seng Chan, Yi-Zhe Song, and Tao Xiang. One loss for all: Deep hashing with a single cosine similarity based learning objective. *Advances in Neural Information Processing Systems*, 34:24286–24298, 2021. 4

[17] Wassily Hoeffding and Herbert Robbins. The central limit theorem for dependent random variables. *Duke Mathematical Journal*, 15(3):773 – 780, 1948. 5

[18] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000. 4

[19] Stefan Klein, Marius Staring, and Josien P. W. Pluim. Evaluation of optimization methods for nonrigid medical image registration using mutual information and b-splines. *IEEE Transactions on Image Processing*, 16:2879–2890, 2007. 4

[20] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16020–16030, 2021. 1

[21] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6231–6241, 2019. 1, 3

[22] Divya Kothandaraman, Tianrui Guan, Xijun Wang, Shuowen Hu, Ming Lin, and Dinesh Manocha. Far: Fourier aerial video recognition. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 657–676, Cham, 2022. Springer Nature Switzerland. 3, 7

[23] Divya Kothandaraman, Ming Lin, and Dinesh Manocha. Differentiable frequency-based disentanglement for aerial video action recognition. *arXiv preprint arXiv:2209.09194*, 2022. 1, 7

[24] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16266–16275, 2021. 2, 3, 6, 11, 14, 15

[25] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018. 2, 6, 7, 11, 13, 14

[26] Jintao Lin, Haodong Duan, Kai Chen, Dahua Lin, and Limin Wang. Ocsampler: Compressing videos to one clip with single-step sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13894–13903, 2022. 1

[27] Frederik Maes, André M. F. Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16:187–198, 1997. 4, 5

[28] Balmukund Mishra, Deepak Garg, Pratik Narang, and Vipul Mishra. Drone-surveillance for search and rescue in natural disaster. *Computer Communications*, 156:1–10, 2020. 3

[29] Hazar Mliki, Fatma Bouhlel, and Mohamed Hammami. Human activity recognition from uav-captured video sequences. *Pattern Recognition (PR)*, 100:107140, 2020. 3

[30] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa M. Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:502–508, 2020. 1

[31] Lichao Mou, Yuansheng Hua, Pu Jin, and Xiao Xiang Zhu. Event and activity recognition in aerial videos using deep neural networks and a new dataset. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 952–955. IEEE, 2020. 3

[32] Kien Nguyen, Clinton Fookes, Sridha Sridharan, Yingli Tian, Xiaoming Liu, Feng Liu, and Arun Ross. The state of aerial surveillance: A survey. *arXiv preprint arXiv:2201.03080*, 2022. 3

[33] Seong Hyeon Park, Jihoon Tack, Byeongho Heo, Jung-Woo Ha, and Jinwoo Shin. K-centered patch sampling for efficient video recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 160–176. Springer, 2022. 7

[34] Asanka G Perera, Yee Wei Law, and Javaan Chahl. Drone-action: An outdoor recorded drone video dataset for action recognition. *Drones*, 3(4):82, 2019. 3

[35] Asanka G Perera, Yee Wei Law, Titilayo T Ogunwa, and Javaan Chahl. A multiviewpoint outdoor dataset for human action recognition. *IEEE Transactions on Human-Machine Systems*, 50(5):405–413, 2020. 3

[36] Josien P. W. Pluim, J. B. Antoine Maintz, and Max A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, 22:986–1004, 2003. 4

[37] Daniel B. Russakoff, Carlo Tomasi, Torsten Rohlfing, and Calvin R. Maurer. Image similarity using mutual information of regions. In *European Conference on Computer Vision*, 2004. 4, 5

[38] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1

[39] Waqas Sultani and Mubarak Shah. Human action recognition in drone videos using a few aerial training examples. *Computer Vision and Image Understanding*, 206:103186, 2021. 3

[40] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2014. 3

[41] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2017. 1

[42] Paul A. Viola and William M. Wells. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24:137–154, 1995. 4, 5

[43] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, 2016. 1, 3

[44] Xijun Wang, Ruiqi Xian, Tianrui Guan, Celso M. de Melo, Stephen M. Nogar, Aniket Bera, and Dinesh Manocha. Aztr: Aerial video action recognition with auto zoom and temporal reasoning. *arXiv preprint arXiv:2303.01589*, 2023. 3, 7

[45] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S. Davis. Adaframe: Adaptive frame selection for fast video recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1278–1287, 2018. 1, 3

[46] Ruiqi Xian, Xijun Wang, and Dinesh Manocha. Mitfas: Mutual information based temporal feature alignment and sampling for aerial video action recognition. *arXiv preprint arXiv:2303.02575*, 2023. 1, 6, 7

[47] Yin-Dong Zheng, Zhaoyang Liu, Tong Lu, and Limin Wang. Dynamic sampling networks for efficient action recognition in videos. *IEEE transactions on image processing*, 29:7970–7983, 2020. 3

[48] Yuan Zhi, Zhan Tong, Limin Wang, and Gangshan Wu. Mgsampler: An explainable sampling strategy for video action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1513–1522, 2021. 2, 3, 6, 7, 11, 12, 13, 14, 15

[49] Yuan Zhi, Zhan Tong, Limin Wang, and Gangshan Wu. Mgsampler: An explainable sampling strategy for video action recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1493–1502, 2021. 2, 3, 11