Video Recognition in Portrait Mode

Mingfei Han[2,1,3,4*] Linjie Yang[1] Xiaojie Jin[1] Jiashi Feng[1] Xiaojun Chang[2,4]
Heng Wang[1]
[1]Bytedance [2]ReLER Lab, AAII, UTS [3]Data61, CSIRO [4]MBZUAI
https://mingfei.info/PMV/

Abstract

The creation of new datasets often presents new challenges for video recognition and can inspire novel ideas while addressing these challenges. While existing datasets mainly comprise landscape mode videos, our paper seeks to introduce portrait mode videos to the research community and highlight the unique challenges associated with this video format. With the growing popularity of smartphones and social media applications, recognizing portrait mode videos is becoming increasingly important. To this end, we have developed the first dataset dedicated to portrait mode video recognition, namely PortraitMode-400. The taxonomy of PortraitMode-400 was constructed in a data-driven manner, comprising 400 fine-grained categories, and rigorous quality assurance was implemented to ensure the accuracy of human annotations. In addition to the new dataset, we conducted a comprehensive analysis of the impact of video format (portrait mode versus landscape mode) on recognition accuracy and spatial bias due to the different formats. Furthermore, we designed extensive experiments to explore key aspects of portrait mode video recognition, including the choice of data augmentation, evaluation procedure, the importance of temporal information, and the role of audio modality. Building on the insights from our experimental results and the introduction of PortraitMode-400, our paper aims to inspire further research efforts in this emerging research direction.

1. Introduction

Most efforts in video recognition have focused on improving the accuracy and efficiency of different models and architectures on public benchmarks. Over the past two decades, there has been a dramatic shift in the types of video recognition models, starting from bags of features [34, 41, 42, 45, 49, 54–56], moving on to convolutional neural networks [8, 13, 14, 22, 33, 52, 53, 58, 59, 65, 66], and more recently, vision transformers [1, 2, 4, 7, 12, 30, 32, 35, 36, 40, 67]. With the evolution of various models, video datasets have played a crucial role in driving each generation of models. The introduction of each video dataset has guided the research community to focus on new challenges. We have moved from using datasets collected in controlled environments (*e.g.*, KTH [43], Weizmann [5]) to more realistic videos (*e.g.*, UCF101 [46], HMDB51 [27]), and now to large-scale web video datasets (*e.g.*, Kinetics-700 [9], HowTo100M [37]).

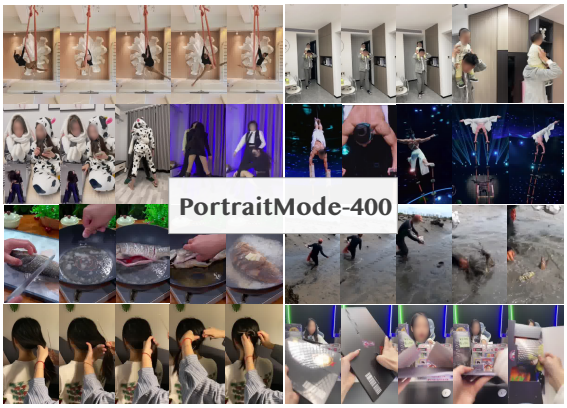


Figure 1. A glance of PortraitMode-400, which is the first dataset dedicated to portrait mode video recognition. It covers videos from 9 domains and 400 specific categories. We show video samples (left to right, top to down) for *aerial yoga*, *riding neck*, *partner dancing (pop music)*, *acrobatics*, *cooking fish soup*, *catching crab*, *styling hair with hairpins* and *opening mystery card packs*, from different domains of our dataset.

While existing video datasets are mostly built on landscape mode videos, portrait mode videos have become increasingly more popular on major social media applications. The shift from landscape mode to portrait mode is not just changing the aspect ratios of the videos. It has significant implications for the types of content that are created and the spatial bias inherent in the data. Portrait mode videos bring in distinct challenges for video recognition as well. For example, they tend to focus more on the subject (*i.e.*, typically humans) with much less background context,

*Work is done during an internship at ByteDance.

and include more egocentric content. In addition, they contain a lot of verbal communication that is essential to understand the video content. There is a pressing need for portrait mode video datasets to explore these new problems.

This paper introduces the first dataset dedicated to portrait mode video recognition, named PortraitMode-400 (abbreviated as PM-400), shown in Figure 1. The dataset consists of 76k videos collected from Douyin[1], a popular short-video application, and annotated with 400 categories. The taxonomy of PM-400 is built in a data-driven way by aggregating search queries and covers a wide range of categories, including sports, food, music, handicrafts, and daily activities, among others. Many of the categories are fine-grained, as shown in Figure 2 (a). The data annotation was performed by professionally trained human annotators, and additional quality assurance was conducted to improve the annotation accuracy and consistency. We built PortraitMode-400 as a single-label dataset, and removed videos that can be tagged with multiple labels during annotation. While the recent 3Massiv [18] dataset also includes a significant percentage of portrait mode videos, it is mostly built for multilingual and multi-modal research, and only has 34 coarse visual concepts, unlike PortraitMode-400.

In addition to introducing the PortraitMode-400 dataset, we have also made preliminary attempts to investigate several critical research problems related to portrait mode video recognition:

- How well does a model trained on landscape mode videos perform on portrait mode videos, and vice versa? We investigate this question by constructing a subset from the Kinetics-700 dataset [9] for a rigorous comparison and visualize classification heatmaps (shown in Figure 3 and Figure 4) to reveal the differences in spatial bias resulting from the change in video format.
- What are the optimal training and testing protocols for portrait mode video recognition? We delve into various components of state-of-the-art deep learning systems, such as data augmentation, evaluation cropping strategies, *etc*. Our discoveries challenge the existing conventions for landscape mode videos, thus necessitating further exploration into portrait mode videos.
- How important is temporal information for portrait mode videos? Can we recognize the actions from single frames [17] or do we need to utilize temporal information for accurate results? We explore different strategies to leverage temporal information and find that temporal information can substantially boost recognition accuracy.
- Audio is another critical modality for video understanding [15, 24]. Does audio help portrait mode video recognition? Our experiments show that even simple audio integration can improve recognition accuracy, indicating the importance of multimodal understanding for portrait mode videos.

[1] Douyin is a popular social media application built for smartphones and primarily features portrait mode short-form videos. https://www.douyin.com/

2. Related Work

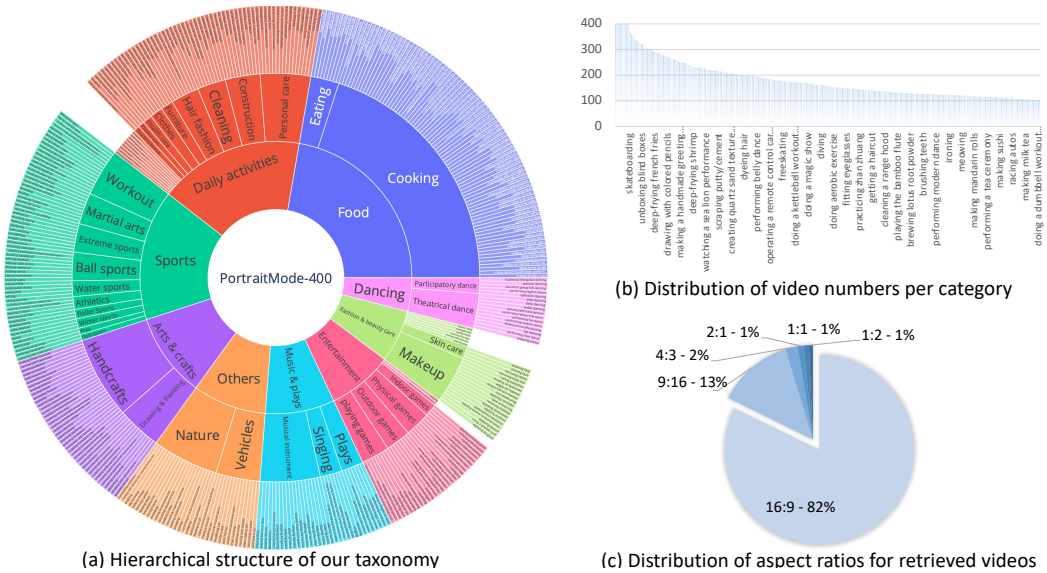
Datasets play a crucial role in investigating research problems, particularly in applications like video recognition. Several pioneering datasets for video recognition were collected in a controlled setting, including KTH [43], Weizmann [5], IXMAS [62], and UIUC [51], *etc*. The videos in these datasets are typically staged with simple and static backgrounds, and human actors are instructed to perform scripted actions repeatedly. By simplifying the action recognition problem, these datasets allow the models to focus on the action of interest. They have inspired the development of hand-crafted features [25, 28, 29, 55] in combination with the bag of features models [11, 45].

Popular video websites, such as YouTube, have become the primary source of video datasets. Unlike the controlled datasets, Internet videos are more realistic and challenging due to factors like background clutter, camera motion, *etc*. Several datasets are created by collecting videos from websites like YouTube, such as UCF101 [46], HMDB51 [27], Activitynet [21], Kinetics-400 [23], Moments in Time [38], *etc*. These datasets serve as the primary testbeds for the development of many successful CNN architectures [8, 13, 14, 22, 33, 52, 53, 58, 59, 65, 66] and vision transformer models [1, 2, 4, 7, 12, 30, 32, 35, 36, 40, 67] in the deep learning era. A recent trend is to build large-scale pre-training datasets, such as HowTo100M [37] and WebVid-10M [3], using text supervision instead of labelled categories.

Social media applications have experienced tremendous growth in recent years, creating a new type of video data known as portrait mode short-form videos. These videos differ significantly from conventional landscape videos used in previous datasets, inspiring us to create a dataset dedicated to portrait mode videos. It is worth noting that the 3Massiv dataset [18] also includes a significant proportion of portrait mode videos. However, it was intentionally designed for multi-lingual and multi-modal purposes, focusing on visual concepts rather than specific actions, with only 34 coarse concepts in total.

3. The PortraitMode-400 dataset

In this section, we provide a comprehensive overview of the process behind constructing our PortraitMode-400 dataset. We begin by discussing our data-driven approach to building a taxonomy, which is based on user queries. Next, we detail our rigorous annotation process and the criteria we applied to ensure high-quality and consistent annotations. Finally, we compare PortraitMode-400 with existing datasets that are relevant to our work, highlighting the unique contributions and advantages of our dataset.



(a) Hierarchical structure of our taxonomy

(b) Distribution of video numbers per category

(c) Distribution of aspect ratios for retrieved videos

Figure 2. Overview of our dataset. (a) We construct our taxonomy in a three-level hierarchical structure, which contains 9 domains and 400 leaf-node categories. (b) We show the distribution of video numbers per category in our dataset, which contains a relatively balanced distribution of categories. (c) We plot the distribution of aspect ratios for the retrieved videos via search queries. The majority of videos (over 85%) are in portrait mode, with 16:9 being the dominant format.

3.1. Taxonomy

The videos in PortraitMode-400 were sourced from Douyin[1]. To better capture the various types of content that portrait mode videos can provide, we created a new taxonomy for PortraitMode-400 instead of reusing categories from existing datasets. Our approach involved building the taxonomy based on popular search queries from Douyin users, which often include text descriptions about the corresponding videos. However, we found that many search queries lacked visual semantic meaning, such as celebrity names or song names. To address this, we manually selected candidate queries containing verbs (*e.g.*, "eating cakes") or nouns indicating potential actions (*e.g.*, "concealer" which often leads to videos about how to use a concealer). After manually examining approximately 38k search queries, we identified about 2.4k usable queries with corresponding videos that might contain actions or motions, as we aimed to incorporate more temporal information.

With the initial set of selected search queries, our second step is to recursively aggregate the queries in a bottom-up manner. This process generates increasingly abstract concepts, resulting in a hierarchical tree structure taxonomy, as illustrated in Figure 2 (a). In addition to producing the final taxonomy, we have two other objectives in this step: 1) merging similar queries into a final leaf node category of the taxonomy; 2) splitting or removing queries that may overlap with existing categories, so that all final categories are mutually exclusive. For example, we merge *tutorials for fitness*, *exercises for weight loss* and *fat-burning fitness exercises* to *aerobics*; we split *calligraphy exercise* into *pen calligraphy* and *brush calligraphy*. After completing the second step, we obtained about 500 candidate categories derived from the 2.4k selected search queries, which are organized in a three-layer hierarchy as depicted in Figure 2.

The taxonomy used in the Kinetics-400 [23] dataset is built through a combination of reusing categories from previous datasets and crowdsourcing. In contrast to Kinetics-400, our taxonomy is developed using a data-driven approach that better reflects the current trends in social media. Besides, our taxonomy covers a wider range of content, including everyday activities (*food*, *beauty care*, *entertainment*, *etc.*) , natural phenomena (*raining*, *snowing*, *etc.*) as well as transportation-related activities (*airplane taking off*, *launching rocket*, *etc.*). This is in contrast to existing datasets that mostly focus on human actions. Furthermore, our taxonomy offers more fine-grained categories compared to 3Massiv [18], which is designed for coarse visual concept classification. For instance, while 3Massiv has only one class for food, our taxonomy includes 89 distinct categories under the food parent node, covering various types of food and food-related activities such as cooking and eating.

3.2. Sampling and annotation

For each of the 500 candidate categories in the taxonomy, we have about 2 to 50 selected search queries associated with it, as described in Section 3.1. We retrieve 1.2k to 740k videos for each query from Douyin[1] depending on how frequently the query has been searched. Subsequently, we create a pool of videos for each category by aggregating all the

Dataset	% of PM	# of Classes	# of Videos	Duration	Avg. Duration	Year
S100-PM [9]	100%	100	20k	1s-10s	9s	'19
3Massiv [18]	95%	34	50k	5s-2min	20s	'21
PortraitMode-400	100%	400	76k	2s-1min	27s	'23

Table 1. Comparison of different portrait mode video datasets. S100-PM is a portrait-mode-only subset sampled from Kinetics-700, as detailed in Section 3.3. 3Massiv contains 5% landscape mode videos and is designed for coarse visual concept recognition. Our PortraitMode-400 contains portrait mode videos only and has more videos in a diversified taxonomy (400 classes).

retrieved videos from their corresponding queries. Figure 2 (c) illustrates the distribution of the aspect ratio of the retrieved videos. Although 16:9 is the dominant aspect ratio, there are also other aspect ratios for portrait mode videos, such as 4:3. For the video pool, we use a few criteria to sample target videos for annotation: 1) we select videos whose aspect ratios (height/width) are greater than 1 to ensure that PortraitMode-400 includes only portrait mode videos; 2) we select videos whose duration is shorter than 1 minute to limit annotation costs; and 3) we select videos that have been viewed over 700 times by Douyin users to ensure that our dataset better reflects the typical types of content for portrait mode videos.

Finally, we perform deduplication on the video pool to eliminate duplicated or similar videos. To achieve this, we extract feature vectors of each video using UniformerBase [30] pretrained on Kinetics-700 dataset [9]. Next, we build a graph by connecting video pairs with feature vectors having a cosine similarity greater than 0.98. We then apply the Louvain algorithm [6] on the graph to identify video clusters and discard all the videos in each cluster except one. About 25% of videos are removed through deduplication, and only videos that meet all the aforementioned criteria move on to the next stage for human annotation.

The human annotation task is straightforward. An annotator is presented with a given category and its video pool, and is asked to confirm or deny whether the category name is a good match for the content of each video. Before starting annotation, annotators undergo training to learn the annotation criteria for all candidate categories, and they are required to pass a quality check test. Only annotators with an accuracy greater than 95% are qualified for annotation to ensure the accuracy and consistency of their annotations. During annotation, annotators discard videos that may be confused with multiple categories of our taxonomy, ensuring that PortraitMode-400 is a strictly single-label dataset. Under our restricted rules, approximately 65% of videos are rejected. To ensure annotation quality, approximately 20% of annotations are reviewed by two additional examiners.

3.3. Comparisons with existing datasets

After finishing annotating all the videos, we keep all the categories that have at least 100 videos. We keep at most 400 videos per category so that the distribution of videos across different categories are more or less balanced, as shown in Figure 2 (b). Our dataset contains 76k videos in total, spanning over 400 categories. We randomly sample 50 videos per category for testing, and the rest are used for training. Table 1 compares the statistics of PortraitMode-400 with other relevant datasets. Though 3Massiv mostly includes portrait mode videos, it is a multi-lingual and multimodal dataset designed for concept recognition with only 34 coarse concepts. PortraitMode-400 has a more diversified and fine-grained taxonomy that is dedicated for portrait mode video recognition.

To conduct a rigorous comparison between landscape mode and portrait mode video recognition, we created two subsets from the Kinetics-700 dataset: a portrait mode subset and a corresponding landscape mode subset. The details of these subsets are shown in Table 1. We first constructed the portrait mode subset, named **Selected-100** Portrait Mode (S100-PM), using the top 100 categories with the most portrait mode videos in Kinetics-700. Each category in S100-PM contains 160 to 352 portrait mode videos, resulting in a total of 20k videos. To build a counterpart landscape mode version from Kinetics-700, we sampled the same number of landscape mode videos as S100-PM for each category, resulting in a landscape mode subset named Selected-100 Landscape Mode (S100-LM). Therefore, S100-PM and S100-LM have the same taxonomy and the same video distribution per category. Although the video content of S100-PM and S100-LM may differ due to different video formats, we believe that they are still useful benchmarks for illustrating and validating the difference between landscape mode and portrait mode video recognition. We have also tried AutoFlip[2] to convert landscape mode videos to portrait mode, thereby ensuring the same video content in both subsets. However, the converted portrait mode videos had unsatisfactory data quality. Thus, building S100-PM and S100-LM from Kinetics-700 remains the best option for rigorously comparing different video formats on recognition tasks.

[2] https://ai.googleblog.com/2020/02/autoflip-open-source-framework-for.html

Model	Train	Val.	Acc.	GFLOPs×views
X3D-M[13]	PM	PM	**52.0**	4.9×3×10
		LM	41.2	4.9×3×10
	LM	PM	**44.5**	4.9×3×10
		LM	43.5	4.9×3×10
Uniformer-S[30]	PM	PM	**42.0**	41.8×1×4
		LM	36.2	41.8×1×4
	LM	PM	40.1	41.8×1×4
		LM	**40.8**	41.8×1×4
MViTv2-S[32]	PM	PM	**41.0**	64.0×1×5
		LM	35.7	64.0×1×5
	LM	PM	33.7	64.0×1×5
		LM	**36.3**	64.0×1×5

Table 2. Cross mode evaluation with different models on Selected-100. Evaluation results performed on the PM subset correspond to the last column of Table 3. Views during inference are shown by the multiplication of # of spatial crops and # of temporal views. Rows highlighted perform best for the corresponding model.

4. Landscape Mode *vs*. Portrait Mode

Landscape and portrait mode videos, often shot in different ways and purposes, display unique content and biases. This affects subjects' action patterns and overall visual dynamics. Therefore, models trained on one mode may struggle in the other. This section examines how models adapt across these different modes, focusing on their spatial information and cross-mode generalizability.

4.1. Cross Mode Evaluation

To show the impact of the different domain priors of landscape and portrait mode videos on video recognition tasks, comparisons need to be made between the same video content shot in portrait mode and landscape mode. Ideally, for each action or event, we should shoot it with both portrait mode and landscape mode cameras. However, such a process is time-consuming and hard to achieve. Therefore, we opt for sampling original portrait mode videos and landscape mode videos with the same distribution and taxonomy from Kinetics-700 [9], as detailed in Section 3.3.

To explore the impact of the different priors to video recognition models, we conducted extensive experiments using different subsets of S100 (S100-PM and S100-LM). We trained various models on different subsets and evaluated their performance on landscape mode videos and portrait mode videos, by randomly selecting 25% videos as the validation set for each subset. For example, evaluated on S100-PM, models trained with S100-PM and S100-LM respectively can be fairly compared to see which video type is more effective to train models for videos in portrait mode. We conduct the experiments on three models, i.e. a CNN model X3D [13], a hybrid transformer model Uniformer [30], and a pure transformer model MViTv2 [32] to show the impact of video formats on different model architectures. During training and testing, we resize frames based on the shorter side while preserving aspect ratios and crop them into 224×224 pixel squares for input. We train all models from scratch without pretraining to avoid the impact of pretraining dataset. Popular pretraining datasets such as ImageNet [26] are biased towards landscape images which may add additional bias to our analysis.

We summarize all results as in Table 2. By comparing results in each row, we find that models trained on PM videos has a larger performance gap on the PM and LM testsets than models trained with LM videos. Moreover, models trained on PM data usually have better performance on PM testset compared to the models trained with LM videos. For example, evaluated on S100-PM, X3D trained with PM videos outperforms the model trained with LM videos by a large margin of 8% (52.0% vs. 44.5%). When evaluated on S100-LM, X3D achieves relatively comparable performance either trained with PM videos or LM videos (41.2% vs. 43.5%). This indicates that training videos in portrait mode are necessary to achieve satisfying performance on portrait mode videos.

4.2. Spatial priors

To investigate the different spatial data priors of portrait mode videos and landscape mode videos, we extensively evaluate the models trained on S100-PM and S100-LM on different frame positions to show the importance of frame features at different locations.

Specifically, we first train Uniformer-S [30] with 112×112 crops and shorter-side resized (set to a random value between 256 and 320) frames on either S100-PM or S100-LM. We name the resulted two models Probing-P and Probing-L. Then we evaluate the models with crops of 112×112 on different locations in a sliding window at the shorter-side resized video clips. The sliding strides vary for portrait mode and landscape mode videos in both height and width. For portrait mode videos, the stride in height is set to 1/16 of the frame height and the stride in width is set to 1/9 of the frame width. Sliding strides of landscape mode videos are adjusted vice versa.

Using Probing-P and Probing-L, we compose an accuracy map of size 16×9 from the accuracies obtained from the different evaluation positions on the S100-PM validation set as shown in Figure 3 (a) and Figure 3 (b). We further compute the difference between the two heat maps in Figure 3 (a) and (b) and obtain the difference map as in Figure 3 (c). Here, the difference value in each position indicates the gap of recognition abilities of the same model trained on landscape mode videos and portrait mode videos, respectively. If a value on the different map is greater than 0, it indicates that Probing-P achieves higher accuracy than Probing-L. For example, as outlined by the yellow boxes in Figure 3 (c), mark 1 indicates the model trained with PM

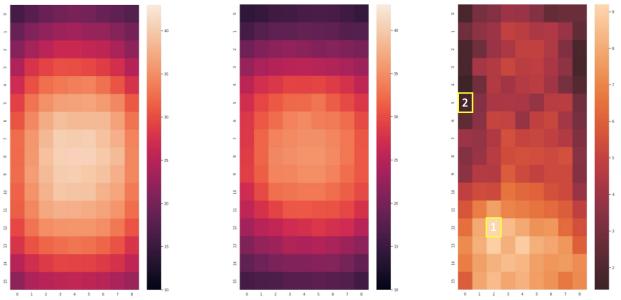


Figure 3. The heatmaps of evaluating the Probing-P (a) and Probing-L (b) at different spatial locations on the validation set of S100-PM. (c) shows the accuracy differences between Probing-P and Probing-L.

videos is stronger to recognize the video categories at this location, while mark 2 indicates models trained by PM and LM videos have similar performance at this location. In general, it can be inferred from the brighter areas in Figure 3 (a) that informative areas in PM videos are more densely concentrated at the middle to lower half of the video. It can also be inferred from Figure 3 (c) that the bottom part of the PM videos contains specific domain knowledge that does not exist in the LM videos, leading to bad performance of models trained on LM videos in this region.

Similarly, we show the accuracy heat maps of the Probing-L and Probing-P evaluated on the LM videos in Figure 4 (a) (b), with the difference of the two heat maps shown in Figure 4 (c). It can be seen that the informative areas in LM videos are in the center part of the video, and the left and right sides on the video frame contain specific domain knowledge that cannot be learned from PM videos. For example, some actions with a wide background in LM videos may not have similar visual cues in the PM videos.

5. Comparison of data preprocessing recipes

Effective data preprocessing is essential for achieving high performance in video classification tasks. In this section, we investigate the impact of different data preprocessing strategies on the performance of portrait mode video recognition. We hypothesize that videos in different aspect ratios may require different crop resolutions for optimal performance. To test this hypothesis, we perform extensive experiments on various portrait mode video datasets, using different crop resolutions and data augmentation techniques. Through our experiments, we identify the best recipes for portrait mode videos when using CNN or transformer models, which are different from that of landscape mode videos.

5.1. Resizing and area sampling

Resizing and cropping are critical steps in the data preprocessing pipeline for video recognition, as they allow videos

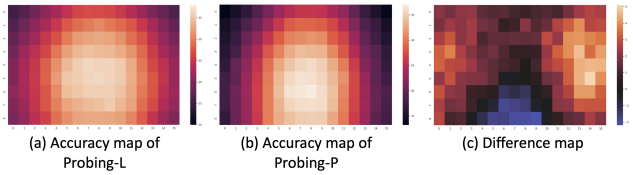


Figure 4. The heatmaps of evaluating the Probing-L (a) and Probing-P (b) at different spatial locations on the validation set of S100-LM. (c) shows the accuracy difference between Probing-L and Probing-P.

to be processed efficiently and are also important ways of data augmentation. Different models in various architectures adopt different strategies. The two popular strategies are the Inception-style method [12, 14, 30, 35, 48, 50], and the shorter-side resizing method [44]. In this subsection, we will explore these two methods in more detail and investigate their effectiveness for portrait mode video recognition.

The shorter-side resizing method is widely used in video recognition methods [4, 8, 10, 13, 31, 39, 47, 53, 57, 59–61, 63, 64, 66, 68]. It involves resizing the video frames so that the shorter side of the frame is set to a length that is fixed [10, 64] or randomly sampled within a range [4, 8, 13, 31, 39, 47, 53, 57, 59–61, 63, 66, 68], while the longer side is scaled proportionally. Then the frames are centre-cropped to a square shape, typically 224×224 and passed into the model. This approach ensures that the input frames have a consistent aspect ratio and are cropped without distortion. In contrast, the Inception-style method augments the shorter-side resizing method with two additional random sampling steps. The first one is to sample a target pixel number from the whole-size video frame by the random ratio between 8% and 100%. Then, it randomly samples an aspect ratio between 3/4 and 4/3 and reshapes the crop area accordingly. Finally, it crops the frames at a random position and resizes them to a fixed resolution in squares (*e.g.*, 224×224) without keeping the aspect ratio. This approach can sample a diverse set of inputs and is designed to adapt the model to videos in different sizes.

We carry out extensive experiments on models of different architectures with the two resizing strategies in Table 3. To alleviate the bias introduced by mixed-orientation data, the models are trained from scratch and we keep any other training setup identical to their original papers, except for learning hyper-parameters, such as batch size and learning rate. During inference, identical augmentation and sampling methods are adopted for different recipes. We guide the readers to supplemental materials for more details.

As shown in Table 3, each model is evaluated on three different portrait mode video benchmarks. For the CNN-based model, *i.e.*, X3D-M [13], the random scaling strategy from the Inception-style method brings an improvement of 2.2% (54.2% vs. 52.0%) on S100-PM [9], 1.1% (53.7%

Model	Data	Incep.	Short.	GFLOPs×views
X3D-M[13]	S100-PM	**54.2**	52.0	4.9×3×10
	3Massiv	**53.7**	52.6	4.9×3×10
	PM-400	**61.7**	61.2	4.9×3×10
Uniformer-S[30]	S100-PM	39.7	**42.0**	41.8×1×4
	3Massiv	42.8	**43.6**	41.8×1×4
	PM-400	50.2	**50.4**	64.0×1×5
MViTv2-S[32]	S100-PM	36.9	**41.0**	64.0×1×5
	3Massiv	50.4	**52.1**	64.0×1×5
	PM-400	61.7	**62.0**	64.0×1×5

Table 3. Comparison of top-1 accuracy (%) of different resizing and area sampling strategies for portrait mode videos, *i.e.*, inception style (Incep.) and shorter-side style(Short.). Views during inference are shown by the multiplication of # of spatial crops and # of temporal views.

Model	Data	Training crops		
		224×224	256×192	288×192
X3D-M[13]	S100-PM	**52.0**	51.6	50.8
	3Massiv	**52.6**	52.5	50.8
	PM-400	**61.2**	61.0	60.8
Uniformer-S[30]	S100-PM	42.0	43.3	**45.4**
	3Massiv	43.6	44.6	**45.8**
	PM-400	50.4	50.8	**51.6**
MViTv2-S[32]	S100-PM	41.0	40.0	**45.5**
	3Massiv	52.1	52.3	**53.8**
	PM-400	62.0	61.4	**62.8**

Table 4. Top-1 accuracy (%) of different training crop resolutions. The models are always tested with the same square crops in **224×224** to ensure the same inference cost across different training crop resolutions.

vs. 52.6%) on 3Massiv [18] and 0.5% on PM-400. Differently, as for the transformer-based models, *i.e.*, Uniformer-S [30] and MViTv2-S [32], randomly scaled input crops bring down the accuracy by a large margin. For example, the random scaling reduces the performance of Uniformer-S by 2.3% (42.0% vs. 39.7%) on S100-PM, 0.8% (43.6% vs. 42.8%) on 3Massiv and 1.3% (72.1% vs. 70.8%) on PM-400. MViTv2-S also shows performance drops from 0.3% to 4.1% across benchmarks. This suggests that optimal strategies diverge from those used in mixed orientation benchmarks like Kinetics[23].

It may be hard to determine the cause of the interesting phenomenon, but we can make a reasonable assumption that it is due to the different data priors in portrait mode only video benchmarks, such as S100-PM and PM-400. With portrait mode videos, the object and its movement are typically limited to a vertical space, which may result in unique visual patterns that are not present in hybrid orientation benchmarks, such as Kinetics. While the cause requires further investigation, these results suggest that there may be unique characteristics of portrait mode videos that require specialized recognition methods.

Model	Data	Testing crops	
		256×192	288×192
X3D-M[13]	S100-PM	$51.4_{0.2\downarrow}$	$50.4_{0.4\downarrow}$
	3Massiv	$52.6_{0.1\uparrow}$	$52.0_{1.2\uparrow}$
	PM-400	$62.9_{1.9\uparrow}$	$63.1_{2.3\uparrow}$
Uniformer-S[30]	S100-PM	$44.4_{1.1\uparrow}$	$46.5_{1.1\uparrow}$
	3Massiv	$45.7_{1.1\uparrow}$	$47.3_{1.5\uparrow}$
	PM-400	$51.9_{1.1\uparrow}$	$53.3_{1.7\uparrow}$
MViTv2-S[32]	S100-PM	$39.8_{0.2\downarrow}$	$46.8_{1.30\uparrow}$
	3Massiv	$52.7_{0.4\uparrow}$	$54.8_{1.0\uparrow}$
	PM-400	$62.1_{0.7\uparrow}$	$63.7_{0.9\uparrow}$

Table 5. Top-1 accuracy (%) of using the same resolution for both training and testing crops. We also report the performance difference compared with using 224×224 testing crops from the first column of Table 4, where ↑ means higher result.

5.2. Shape of frame crop

In this subsection, we explore the impact of different crop strategies on model performance in portrait mode video recognition. Specifically, we investigate the performance of models trained and tested on crops of varying sizes and aspect ratios.

Traditional methods typically use square frame crops to ensure even coverage of object and movement in both vertical and horizontal directions. However, we argue that this approach may not be optimal for portrait mode videos, which typically contain object and movement information in vertical directions. Cropping the frames into squares could potentially result in a loss of critical information and more background noise. As shown in Figure 3, portrait mode videos possess more informative content distributed vertically, and cropping into squares may not effectively capture this information.

To comply with the unique information distributive characteristics, we propose to crop the areas in vertical rectangles and input them directly into models without distortion. We experiment with crops in different aspect ratios and in similar pixel numbers to the square input, *i.e.*, 256×192 and 288×192, in order to fairly compare the models under different input resolutions. With input shape changed, we only modify the last global pooling layer. We keep any other training details identical to the setup using square inputs.

As shown in Table 4, we train models with different input crops on portrait mode video benchmarks and test with square crops, *i.e.*, 224×224 to ensure identical inference cost. It is thrilled to see that increase in aspect ratio introduces continuing performance improvement for transformer-based models, *i.e.*, Uniformer-S and MViTv2-S. We also observe that change in aspect ratio degrades the performance of X3D-M, showing different behaviour to transformer-based models. The potential reason could be due to the fixed square receptive field of convolution networks regardless of the input resolutions, which is not compatible with the elongated image shape.

Data	Model	# of Frames	Top1-Acc.
K400 [23]	Uniformer-frames	16×4	72.1
	Uniformer-S [30]	16×4	$76.6_{4.5\uparrow}$
3Massiv [18]	Uniformer-frames	16×4	41.9
	Uniformer-S [30]	16×4	$42.8_{0.9\uparrow}$
PM-400	Uniformer-frames	16×4	45.7
	Uniformer-S [30]	16×4	$50.3_{4.6\uparrow}$

Table 6. **Temporal information importance**: Effect of utilizing temporal information for video recognition on different benchmarks.

Data	Modality	Top1-Acc.
3Massiv [18]	Visual	52.7
	Audio	31.6
	Visual+Audio	54.9
PM-400	Visual	54.6
	Audio	15.2
	Visual+Audio	57.0

Table 7. **Audio importance**: Comparison of different modalities with offline feature embeddings.

In order to further validate the benefits of rectangular input, we evaluate the performance of X3D-M [13], Uniformer-S [30] and MViTv2-S [32] on non-square training resolutions and tested them on three portrait mode video benchmarks. We find that the three models achieve higher accuracies on 3Massive and PM-400 with both crops in 256×192 and 288×192. On S100-PM, Uniformer-S and MViTv2-S achieve better testing results with 288 × 192 resolution, with FLOPs increased by around 15% (47.5G vs 41.8 for Uniformer-S; 72.7G vs. 64.5G for MViTv2-S). Note that FLOPs of 256×192 are smaller than square 224×224 (single clip inference cost: 40.6G vs. 41.8G for Uniformer-S; 62.9G vs. 64.5G for MViTv2-S). The performance boost further supports the potential benefits of rectangular input for video recognition in portrait mode.

6. The importance of temporal information

In this subsection, we investigate the importance of utilizing temporal information for portrait mode video recognition. We show that the PortraitMode-400 is a valuable resource for evaluating video models in the challenging setting of portrait mode video recognition.

We design two baselines with different temporal utilization approaches and extensively evaluate the models trained on Kinetics-400 [23], 3Massiv [18] and our PortraitMode-400. Specifically, we build our baselines with Uniformer-S and train the models with 224 × 224 crops. Uniformer-frames is constructed with image-based Uniformer-S and temporal aggregation of predicted logits using mean pool. It serves as a naive baseline since the temporal information is incorporated simply by merging the predicted logits across frames. For more advanced temporal correspondance, we train a video-based Uniformer-S endowed with self-attention on temporal dimension, building and learning temporal relations in different levels.

As shown in Table 6, by leveraging temporal self-attention, Uniformer-S obtain accuracy improvement by 4.5% and 4.6% on Kinetics-400 and PortraitMode-400 respectively. Interestingly, the 3Massiv dataset, most of which videos are in portrait mode, does not show as large of a performance gain from using temporal information as our PM-400. In contrast, our PortraitMode-400 dataset shows a significant performance gain from using temporal information, attributable to its diverse collection of videos rich in intricate temporal dynamics.

7. The importance of the audio modality

In this section, we aim to explore the significance of audio information in portrait mode video recognition. To achieve this, we adopt the R3D-50 [19] backbone trained on Kinetics-700 [9] for spatio-temporal modeling and the VGG [20] model trained for sound classification [16] for audio modeling, following the practice in 3Massiv [18]. We freeze the audio-visual backbones and train the classifier and multimodal fusion layers.

Our findings, as presented in Table 7, reveal that the model trained with audio consistently outperforms the model trained without audio on both the PM-400 and 3Massiv by approximately 2.4 points. This indicates that audio information plays a crucial role in portrait mode video recognition. Incorporating audio information can significantly enhance the performance of the model. We argue that audio cues can provide additional information about the subject's actions, emotions, and the surrounding environment, which poses unique challenges for video recognition in portrait mode.

8. Discussions

In this work, we advocate conducting research on portrait mode videos. To this end, we introduce the PortraitMode-400 dataset dedicated for portrait mode video recognition with a fine-grained taxonomy. We also make initial attempts to explore the specific properties of portrait mode videos, including their spatial bias, and the optimal training and evaluation protocols, with effects of the temporal information and audio modality. We believe our dataset can serve as a testbed to facilitate further research such as novel architecture designs and multi-modality modeling on portrait mode videos.

Acknowledgement

This work is partially supported by Australian Research Council (ARC) Discovery Program under Grant No. DP240100181.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6816–6826, 2021. 1, 2

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 1, 2

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 2

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? 2021. 1, 2, 6

[5] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *10th IEEE International Conference on Computer Vision (ICCV 2005), 17-20 October 2005, Beijing, China*, pages 1395–1402, 2005. 1, 2

[6] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008. 4

[7] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. *Advances in Neural Information Processing Systems*, 34:19594–19607, 2021. 1, 2

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 1, 2, 6

[9] João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *CoRR*, abs/1907.06987, 2019. 1, 2, 4, 5, 6, 8

[10] Chun-Fu Richard Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6165–6175, 2021. 6

[11] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, pages 1–2. Prague, 2004. 2

[12] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *arXiv preprint arXiv:2104.11227*, 2021. 1, 2, 6

[13] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 1, 2, 5, 6, 7, 8

[14] 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. 1, 2, 6

[15] 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. 2

[16] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 776–780. IEEE Press, 2017. 8

[17] Guodong Guo and Alice Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10): 3343–3361, 2014. 2

[18] Vikram Gupta, Trisha Mittal, Puneet Mathur, Vaibhav Mishra, Mayank Maheshwari, Aniket Bera, Debdoot Mukherjee, and Dinesh Manocha. 3massiv: Multilingual, multimodal and multi-aspect dataset of social media short videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21064–21075, 2022. 2, 3, 4, 7, 8

[19] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 8

[20] David Harwath, Galen Chuang, and James Glass. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech, 2018. 8

[21] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 961–970. IEEE, 2015. 2

[22] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1, 2

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2, 3, 7, 8

[24] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 2

[25] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008. 2

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural net-

works. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 5

[27] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011. 1, 2

[28] Ivan Laptev and Tony Lindeberg. On space-time interest points. *International journal of computer vision*, 64(2-3): 107–124, 2005. 2

[29] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[30] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition, 2022. 1, 2, 4, 5, 6, 7, 8

[31] Xianhang Li, Yali Wang, Zhipeng Zhou, and Yu Qiao. Smallbignet: Integrating core and contextual views for video classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1092–1101, 2020. 6

[32] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 1, 2, 5, 7, 8

[33] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019. 1, 2

[34] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang Wang. Knowledge distillation via the target-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10915–10924, 2022. 1

[35] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021. 1, 2, 6

[36] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 1, 2

[37] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 1, 2

[38] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 2

[39] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3163–3172, 2021. 6

[40] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021. 1, 2

[41] Xiaojiang Peng, LiMin Wang, Zhuowei Cai, Yu Qiao, and Qiang Peng. Hybrid super vector with improved dense trajectories for action recognition. In *ICCV Workshops*, pages 109–125, 2013. 1

[42] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, 150:109–125, 2016. 1

[43] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, August 23-26, 2004*, pages 32–36, 2004. 1, 2

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[45] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France*, pages 1470–1477, 2003. 1, 2

[46] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1, 2

[47] Ximeng Sun, Rameswar Panda, Chun-Fu Richard Chen, Aude Oliva, Rogerio Feris, and Kate Saenko. Dynamic network quantization for efficient video inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7375–7385, 2021. 6

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 6

[49] Kevin Tang, Bangpeng Yao, Li Fei-Fei, and Daphne Koller. Combining the right features for complex event recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2696–2703, 2013. 1

[50] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 6

[51] Du Tran and Alexander Sorokin. Human activity recognition with metric learning. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille,*

France, October 12-18, 2008, Proceedings, Part I 10, pages 548–561. Springer, 2008. 2

[52] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. 1, 2

[53] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019. 1, 2, 6

[54] Muhammad Muneeb Ullah, Sobhan Naderi Parizi, and Ivan Laptev. Improving bag-of-features action recognition with non-local cues. In *BMVC*, pages 95–1. Citeseer, 2010. 1

[55] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013. 2

[56] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103:60–79, 2013. 1

[57] Heng Wang, Du Tran, Lorenzo Torresani, and Matt Feiszli. Video modeling with correlation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 352–361, 2020. 6

[58] 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 *ECCV*, pages 20–36. Springer, 2016. 1, 2

[59] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1895–1904, 2021. 1, 2, 6

[60] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018.

[61] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 6

[62] Daniel Weinland, Edmond Boyer, and Remi Ronfard. Action recognition from arbitrary views using 3d exemplars. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–7. IEEE, 2007. 2

[63] Chao-Yuan Wu, Ross Girshick, Kaiming He, Christoph Feichtenhofer, and Philipp Krahenbuhl. A multigrid method for efficiently training video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 153–162, 2020. 6

[64] Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Yuille, and Yingwei Li. Learning from temporal gradient for semi-supervised action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3252–3262, 2022. 6

[65] Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. A discriminative CNN video representation for event detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1798–1807, 2015. 1, 2

[66] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 591–600, 2020. 1, 2, 6

[67] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 1, 2

[68] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16804–16815, 2022. 6