Unifying Video Self-Supervised Learning across Families of Tasks: A Survey

Ishan Dave *, Malitha Gunawardhana, Limalka Sadith, Honglu Zhou, Liel David, Daniel Harari, Mubarak Shah, Muhammad Khan

Posted Date: 2 August 2024

doi: 10.20944/preprints202408.0133.v1

Keywords: Video Understanding; Self-Supervised Learning; Representation Learning

 doi:10.20944/preprints202408.0133.v1

Article

Unifying Video Self-Supervised Learning across Families of Tasks: A Survey

Ishan Dave [1,*,†], Malitha Gunawardhana [2,†], Limalka Sadith [3], Honglu Zhou [4], Liel David [5], Daniel Harari [5], Mubarak Shah [1] and Muhammad Haris Khan [6]

[1] University of Central Florida, USA
[2] University of Auckland, New Zealand
[3] University of Moratuwa, Sri Lanka
[4] Salesforce Research, USA
[5] Weizmann Institute of Science, Israel
[6] Mohamed bin Zayed University of Artificial Intelligence, UAE
* Correspondence: ishandave@ucf.edu
† These authors contributed equally to this work.

Abstract: Video self-supervised learning (VideoSSL) offers significant potential for reducing annotation costs and enhancing a wide range of downstream tasks in video understanding. The ultimate goal of VideoSSL is to achieve human-level video intelligence across a spectrum of tasks, from low-level tasks such as pixel temporal correspondence to high-level complex spatio-temporal tasks like action recognition. However, most existing VideoSSL methods focus on isolated aspects of this spectrum and fail to integrate different levels of task complexity. Our study presents the first comprehensive survey that connects all families of VideoSSL methods. We provide a detailed review of the full spectrum of VideoSSL, from low to high levels, by conceptually linking their self-supervised learning objectives and including a comprehensive categorization. Our extensive evaluation highlights the strengths and limitations of each SSL objective across various downstream task families. We also detail the challenges in current VideoSSL research such as data curation, interpretability, deployment, and privacy concerns, an area that previous surveys have not thoroughly explored. In addressing these challenges, we recognize the strengths of existing methods in addressing these challenges and outline future directions for research.

Keywords: video understanding; self-supervised learning; representation learning

1. Introduction

Deep learning methods have significantly advanced the field of video understanding, encompassing tasks such as action recognition, video retrieval, video object segmentation, gait recognition, etc. These advancements have had a profound impact on a wide range of applications, including surveillance, sports analytics, surgical video analysis, content recommendation, and behavioral studies. Through the utilization of deep learning techniques, particularly powerful video architectures like 3D-CNNs [1–4] and video transformer models [5–8] , video understanding systems have gained the ability to accurately analyze and comprehend the complex spatial and temporal dynamics present in videos.

Although powerful video architectures are capable of capturing the intricate dynamics of videos, their effectiveness often relies on the availability of large-scale labeled video datasets, such as Kinetics [3], HACS [9], LSHVU [10], which consist of hundreds of thousands of well-curated labeled videos. However, in practice, annotating videos is a time-consuming, tedious, and expensive process, making it challenging to acquire extensive labeled video data, especially when compared to image datasets. Despite this challenge, there is an immense amount of unlabeled video data readily available through the internet, web platforms, and other sources. Leveraging this vast corpus of unlabeled data can unlock the potential of video understanding and unshackle progress in the field. As a result, label-efficient training paradigms, such as self-supervised learning (SSL), become particularly more pressing in the video domain than in the image domain.

Furthermore, SSL-based approaches can be easily adapted to new domains and downstream tasks, making them an attractive option for real-world video understanding applications. Although video

 doi:10.20944/preprints202408.0133.v1

self-supervised learning (videoSSL) aims to mimic human learning from unlabeled data and generalize across various high-level and low-level downstream tasks, there are limitations. For example, an untrained human observing a 'baseball pitching' action understands it at multiple levels: tracking the athlete's body joints throughout the video (a 'low-level' task), understanding the transition from throw-stance to ball-release (an 'intermediate-level' task), and comprehending the overall action at the video level (a 'high-level' task). However, existing videoSSL literature typically focuses on just one aspect of video understanding, with only a few studies addressing multiple aspects but showing dominant performance in only one. Furthermore, there is no comprehensive survey that connects different videoSSL objectives at a conceptual level.

In this paper, we extensively review and connect the different families of SSL objectives as shown in Figure 1 and limit our scope to the single visual modality only . To the best of our knowledge, there is only one prior survey [11] on videoSSL, which partly covers the field, focusing solely on action-related downstream tasks. Also, the survey [11] also categorizes videoSSL objectives at the surface level: pretext task-based, contrastive learning-based, and generative-based. In contrast, our paper offers a more comprehensive categorization, covering more papers across all families of videoSSL and connecting them conceptually. We categorize videoSSL objectives based on properties such as learning temporal-ordinal information, temporal coherence, spatio-temporal continuity, and instance discrimination. Our extensive set of experimental studies provides insights into which properties aid specific downstream tasks, marking the first attempt to connect various videoSSL methods across different families of downstream tasks.

In addition to providing an extensive survey that conceptualizes connections between various videoSSL families, we also detail the upcoming challenges and potential approaches, aspects missing in prior surveys. We categorize challenges into three areas: (1) Data-related challenges, such as the use of uncurated datasets and untrimmed videos. (2) Deployment challenges, such as robustness to distribution shifts, adversarial attacks, and computational/storage efficiency. (3) Emerging challenges, such as privacy preservation, egocentric videos, and the integration of multiple videoSSL families. Our main contributions can be summarized as follows:

- We provide the first holistic study to cover and connect various families of videoSSL methods comprehensively.
- Extensive evaluations on various downstream tasks using different protocols are presented, offering insights into the performance and utility of learned videoSSL models across different task families.
- We detail various challenges associated with videoSSL methods and recognize the strengths of existing papers that have laid initial foundations in this direction.

 doi:10.20944/preprints202408.0133.v1

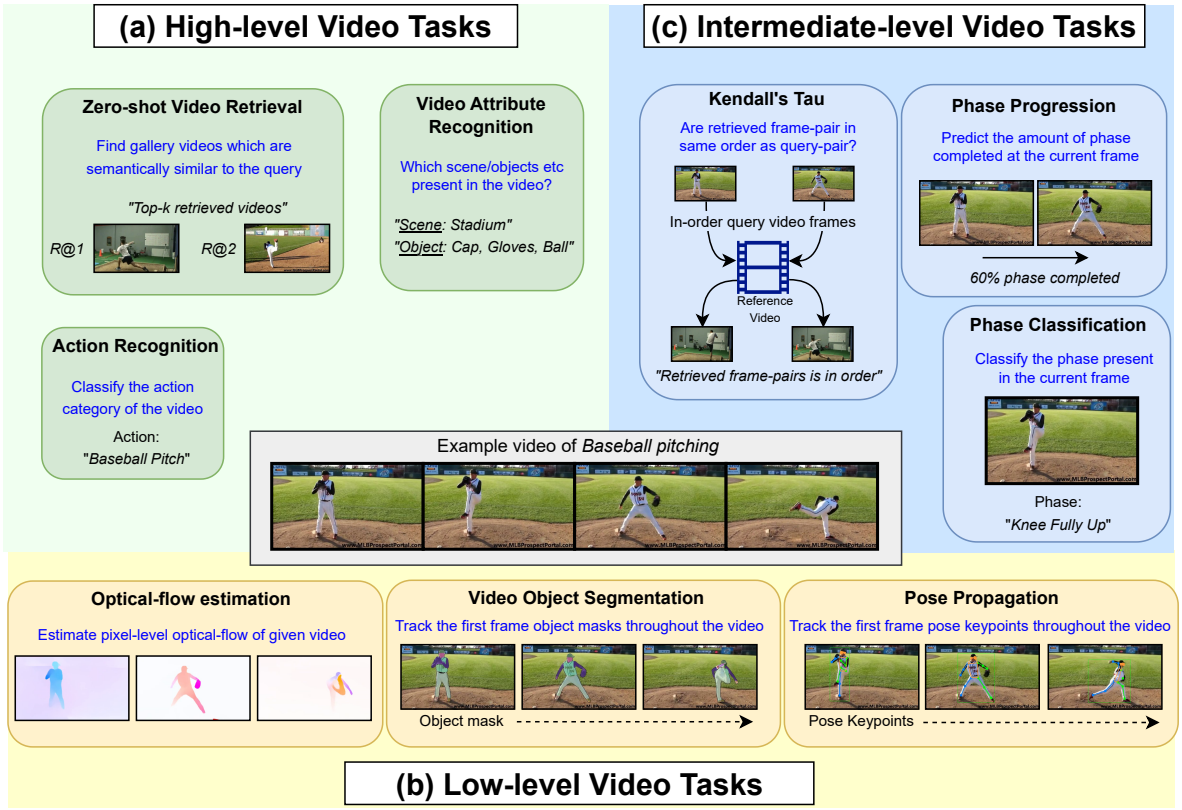


Figure 1. Families of VideoSSL methods: Based on our review of the literature, we identify three major families of videoSSL methods: (a) Methods that focus on high-level semantic tasks which require complex spatio-temporal understanding, such as action recognition, video retrieval, and video attribute classification. (b) Methods that concentrate on low-level video dynamics, primarily learning good temporal correspondences between video segments. Tasks in this category include video object segmentation and pose tracking. (c) Objectives that aim to learn the action-class agnostic internal structure of an action, which falls between high-level semantic understanding and low-level correspondence. These methods focus on identifying frame-level key events and action phases, useful for fine-grained action understanding and temporal alignment of videos. In this paper, we thoroughly review and connect the various families of SSL objectives through extensive evaluations. To the best of our knowledge, only one previous survey [11] on videoSSL exists, which partially addresses the field, concentrating exclusively on action-related downstream tasks.

2. Problem Definition: Video Self-Supervised Learning

The goal of a self-supervised learning method is to first train from unlabeled data (i.e., videos) by optimizing on a self-supervised objective. Once the model is trained, it is evaluated for various downstream tasks such as action recognition, video object segmentation, etc.

SSL Pretraining Phase: During this stage, the model learns to identify and understand the underlying patterns and representations within the unlabeled videos. Let $\mathbb{D}_{\text{unlabeled}} = \{x_1, x_2, \ldots, x_n\}$ represent the set of unlabeled video samples. The goal is to learn weights f_θ by optimizing a self-supervised loss $\mathcal{L}_{\text{SSL}}$, such as contrastive loss or masked reconstruction loss.

$$\theta^* = \arg\min_{\theta} \forall_{x \in \mathbb{D}_{\text{unlabeled}}} [\mathcal{L}_{\text{SSL}}(f_\theta(x))] \tag{1}$$

This approach allows the model to develop a deep understanding of the intrinsic characteristics and complexities of the data without relying on predefined labels or annotations.

Downstream Task Phase: Once the model is trained through the self-supervised objective on $\mathbb{D}_{\text{unlabeled}}$, it is evaluated on various downstream tasks. The downstream tasks may require an additional tuning of the f_θ or additional trainable parameters f_ϕ. Currently, there are four well-known

 doi:10.20944/preprints202408.0133.v1

settings in terms of the tuning requirement in the downstream phase as shown in Figure 2. Some of the video search-based downstream applications, such as zero-shot video-to-video retrieval and video object segmentation, require using the model directly after the SSL pretraining phase, as shown in Figure 2(a). Whereas, the semantic recognition-based downstream tasks such as action recognition could be performed in three different tunable settings: one could be just training a linear classifier layer on top of the frozen model (Figure 2(b)). However, recent methods [12] suggest utilizing attention pooling-based tuning on top of the unpooled SSL features, which is helpful for masking-based self-supervised objectives.

Let $\mathbb{D}_{\text{labeled}} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ represent the set of labeled data samples, where y_i denotes the label for sample x_i. The tunable parameters ϕ are learned on top of the frozen features of the SSL pre-trained model f_{θ^*} by optimizing a supervised loss $\mathcal{L}_{\text{sup}}$, such as cross-entropy loss, as shown in Equation (2). In the case of full-finetuning, Equation (2) is optimized for both ϕ and model weights θ.

$$\phi^* = \arg\min_{\phi} \forall_{(x_i, y_i) \in \mathbb{D}_{\text{labeled}}} [\mathcal{L}_{\text{sup}}(f_\phi(f_{\theta^*}(x_i)), y_i)] \tag{2}$$

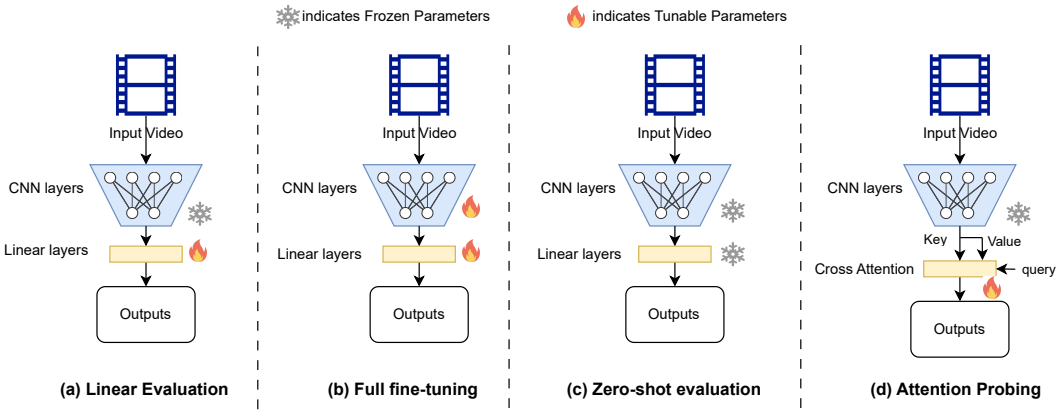


Figure 2. Configurations of Tunable Parameters in the Downstream Task Phase

3. Representation Learning

In our comprehensive survey, we refine the categorization of self-supervised learning objectives in the video domain to present a more detailed and intuitive taxonomy of representation learning. As depicted in Figure 3, our taxonomy organizes the learning objectives into distinct categories, focusing on the intrinsic cues they exploit, whether they be low-level, high-level, precomputed visual priors, or multimodal representations. This structured approach not only clarifies the scope of each category but also highlights the diverse methodologies and their specific contributions to enhancing videoSSL capabilities.

 doi:10.20944/preprints202408.0133.v1

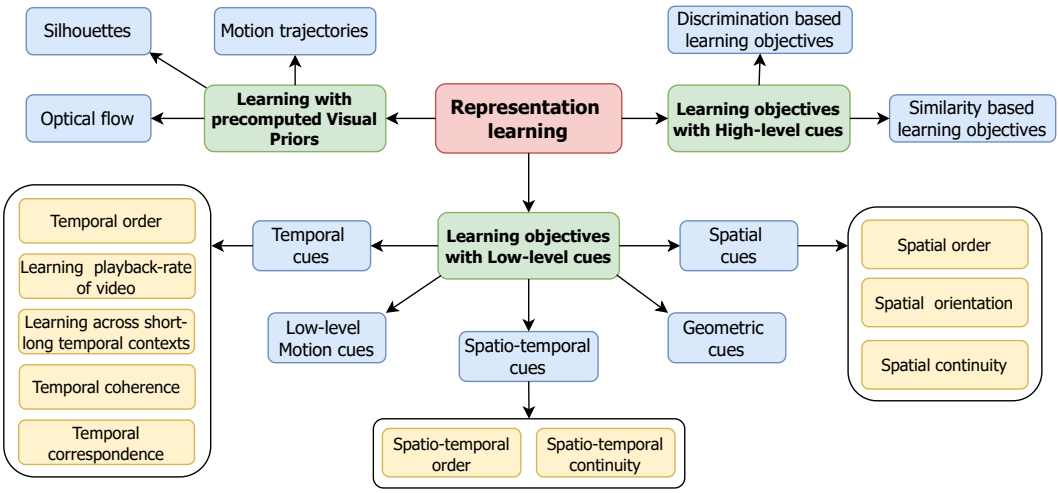


Figure 3. Categorization of the self-supervised objectives We categorize all families of videoSSL based on their self-supervised objectives. We also provide conceptual connection between various self-supervised objectives.

3.1. Learning Objectives with Low-Level Cues

Low-level cues in video self-supervised learning are intrinsic properties extracted directly from raw pixels that provide a foundational understanding of visual content without any semantic interpretation. They consist of the elemental signals such as the subtle changes in pixel intensity that inform about motion, the variance in temporal sequences that reveal playback rates, and the spatial configuration of pixels that elucidate object orientation and scene geometry. These cues underpin tasks such as detecting how an object moves across a sequence of frames (motion detection), recognizing the natural rhythm in a video (playback speed analysis), and inferring the continuity of action (frame smoothness) without relying on higher-level human-assigned labels or annotations. The primary benefit of leveraging low-level cues lies in their intuitiveness and interpretability, clearly delineating which objectives contribute to the learned spatio-temporal representations.

3.1.1. Learning the Temporal Cues

We consider various temporal cues such as temporal order, playback rate, temporal granularity, smoothness, and entity correspondence in the learned representations across the temporal dimension.

Temporal Order

Temporal order in videos refers to the sequence or arrangement of events, actions, or frames over time. In video analysis, understanding the temporal order is crucial for interpreting the sequence of activities or the evolution of scenes in a sensible manner. When using temporal order, the model is required to identify whether the frames are placed in the correct temporal order or not. To achieve that, the model needs to keep track of the temporal dynamics of the moving entity across frames, and by doing so, the model learns rich representations.

In the early stage of videoSSL, Misra *et al.* [13] proposed a novel method called 'Shuffle and Learn'. Let us assume that a video V consists of n number of frames. They take 5 frames out of these n frames and create three tuples (one positive and two negative tuples). Let us assume that these five sampled frames are $(f_p, f_q, f_r, f_s, f_t)$. A positive tuple would be (f_q, f_r, f_s) or (f_s, f_r, f_q) by considering directional ambiguity. Negative tuples would be (f_q, f_p, f_s) and (f_q, f_t, f_s). Now, the problem becomes a classification task using these three types of tuples. They sample these tuples from a high-motion window. Suppose the distance between q and s, which they identify as the difficulty of positives, is higher. In that case, it is harder to identify correspondence across positive pairs and

 doi:10.20944/preprints202408.0133.v1

the minimum distance between p, q and d, e is used as the difficulty of negatives where high value makes them easier. Building upon this concept, the O3N [14] framework takes a further step. Known as the Odd One Out network, this architecture involves identifying the incorrect (odd) frame sequence from multiple clips. Out of $N + 1$ clips, N clips are in the correct temporal order, and one clip has framed shuffled. The O3N network attempts to identify the location of the odd video. In a parallel development, Lee *et al.* [15] proposed a different approach, the Order Prediction Network (OPN), which treats the predicting order as a sequence sorting task. The input of the OPN consists of four randomly shuffled frames, and they group both forward and backward permutations into the same class. The OPN involves two main stages: data sampling, where tuples are selected based on motion magnitude and processed through spatial jittering and channel splitting to emphasize semantics over low-level features, and order prediction, which employs a Convolutional Neural Network (CNN) for feature extraction. The network encodes features for each frame and then generates pairwise features for frame pairs, ultimately predicting their order. Further improving the concept of temporal ordering, AoT [16] uses the arrow of time as a pretext task to predict whether the video is going forward or backwards. Their temporal activation map uses T groups, which contain optical flow frames. However, this network needs data preparation, such as removing black frames in videos and stabilising the camera.

However, both Shuffle and Learn [13] and Odd One Out [14] methods use the order of frames to learn representations using 2D CNNs. In that case, the model is required to understand whether the frames are in order or not. Consider the task of catching a ball. It is tough to identify whether it is throwing or catching using the shuffled frames because both directions are possible. To address this issue, Xu *et al.* [17] propose a pretext task of predicting the order of clips instead of frames using 3D CNNs. Furthermore, Xue *et al.* [18] also propose a new temporal pretext task where they first form a global clip by taking some out-of-order clips in between the in-order clips, and their goal is to find the location of the out-of-order frame in the global clip. Such localization problem increases the challenge of the temporal-order-based pretext tasks and helps in improving downstream performance.

Further advancing this field, the Video Cloze Procedure (VCP) method by Luo *et al.* [19] introduced an innovative technique involving multiple operations, including temporal remote shuffling. In this method, a selected clip from a video sequence is removed and replaced with another clip from a significantly different time point. This technique leverages the similarity in background across temporally spaced frames, focusing the model's attention on the more dynamic and informative foreground elements, thereby enhancing the model's ability to understand and predict the sequence of events in videos.

To improve the representation learning using temporal ordering, later SSL video models propose improvements to classical temporal ordering. Hu *et al.* [20] introduce the Contrast-and-Order Representation (CORP) approach, enabling it to discern both the visual details in each frame and the time sequence across frames. Specifically, their method first determines if two video clips are from the same source. If they are, it predicts their temporal order. By doing so, the model can understand temporal relationships in videos rather than merely comparing two modified clips from a single video without considering their sequence. Also, the SeCo framework [21], employs a temporal order validation task. This task serves as a supervisory signal for video sequences, emphasizing the importance of understanding the inherent temporal order within video content. Similarly, Guo *et al.* [22] propose to use a version of Edit Distance to measure the degree of temporal difference between a video clip and its shuffled version. In addition to these approaches, the research by Luo *et al.* [23] explores the utilization of temporal disordered patterns and [24] leverages a variety of pretext tasks, such as predicting the direction of time to facilitate self-supervised learning of video representations. Furthermore, SCVRL [25] introduces a novel objective where a video clip is compared to the same clip with its temporal order shuffled. This approach ensures that the learned representation acknowledges the temporal sequence of actions, enabling it to distinguish its different phases.

 doi:10.20944/preprints202408.0133.v1

Apart from this, some works utilize the temporal order along with the contrastive learning objective for example TaCo [26] identifies if the sequence is shuffled or in a correct temporal order. The order information is also used along with the graph-based learning in TCGL [27] to predict the snippet order. In TEC [28], the task is to encourage the equivariance along the temporal dimension in the learned representation utilizing the temporal ordering task. Their pretext task is to identify the order of the video pair which could be temporally overlapping or non-overlapping.

Learning Playback-Rate of Video

The playback rate of a video refers to the speed at which the video frames are displayed, typically measured in frames per second (fps). By altering the playback rate, we can change the perceived speed of motion within the video. Playback rate can be changed by skipping different amount of frames in between the two sampled frames. In the context of self-supervised learning for video understanding, by training a model to accurately predict the playback rate, it is forced to learn about the temporal dynamics and motion patterns inherent in the video data. This learning process encourages the model to develop a deeper understanding of the temporal relationships between frames, which is crucial for tasks such as action recognition and event detection. The intuition behind using playback rate prediction as a pretext task is that it requires the model to capture the nuances of motion and temporal changes in the video. For instance, a model that can accurately predict the playback rate of a video of a bouncing ball must understand the physics of the ball's motion, including its acceleration and deceleration as it bounces.

To this end, Cho *et al.* [29] introduced PSPNet which focuses on predicting the order of various speeds and directions in videos. By utilizing clips played back at variable speeds, their network learns to discern and predict the correct playback speed, thereby gaining insights into the temporal dynamics of the videos. In the 'video-pace model' introduced by Wang *et al.* [30], the network is trained using pace-varying video clips, with the objective being to identify the varying paces of these clips. The paces are randomly selected from a set of candidates and the model incorporates two contrastive learning strategies to regularize the learning process in the latent space, simultaneously optimizing both classification and contrastive components for effective training. Rather than predicting the speed of a video, the SpeedNet [31] model predicts the speediness of videos. Speediness is not the same as the magnitude of the motion. In this task, the model is trained for a simple binary classification task which is to identify whether the input videos are at their normal speed or not. However, speeding the videos does not always guarantee that they contain abnormal dynamics. Take the example of walking or running. When speeding up the video, it can be fast walking or fast running which might not be unusual. Building upon Speednet, RSPNet [32] utilizes relative playback speed as their pretext task and naming their self-supervised objective. The primary distinction of RSPNet is its emphasis on detecting the relative speed at which different clips are played. They use instance discrimination tasks to pull same-speed videos together. Advancing in this line of research, Dave *et al.* [33] propose a more complex frame-level time-varying skiprate prediction task (TSP). In contrast to prior work that focuses on identifying the sequence-level skip rate prediction, their approach formulates a sequence with varying skip rates. TSP pretext task involves a more dense prediction, i.e., frame-level prediction between each consecutive sampled frame. They demonstrate that such frame-level tasks encourage improvements in performance over the conventional clip-level skip rate prediction. Apart from these methods, Jenni *et al.* [34] discuss how objects disclose their shape, behavior, and interaction with other objects when in motion where the challenge lies in extracting this information. The study advocates recognizing different types of temporal transformations, especially playback rate, based on the premise that recognizing a time transformation necessitates an accurate model of the inherent video dynamics while [24] uses multiple pretext tasks including speed prediction to learn video representations in a self-supervised manner.

 doi:10.20944/preprints202408.0133.v1

Learning across Short-Long Temporal Contexts

In the video understanding, given a limited number of frames in sampled clip, one can either do sparse sampling and obtain a clip with longer temporal span (more temporal context) or perform dense sampling to obtain temporally rich shorter temporal context. To this end, video self-supervised learning methods have been proposed to facilitate learning across various temporal context, accommodating varying frame rates and global and local perspectives within videos. These approaches are designed to enhance the extraction of beneficial features by leveraging the intrinsic structure of data at different temporal scales of detail.

One of the earlier approaches in learning across the temporal context of local and global clips is from Yang *et al.* [35], where they maximize the mutual similarity between the sparse (i.e. fast) and dense (i.e. slow) video streams. They hierarchically apply such SSL objective at different layers of the model. In the same line, LSFD [36] utilizes long and short videos to encourage learning both stationary and non-stationary video features utilizing the contrastive learning-based objective. They utilize long and short clips from the video where they define stationary features as those consistent across both views, while non-stationary features are compiled from shorter sequences to match the longer sequence they originate from. Similarly, in BraVe [37], one view has access to a narrow temporal window of the video, while the other has broader access to the entire video content. Through BraVe, models are trained to generalize from the narrow temporal view to understand the broader content of the video. Another work MaMiCo [38] aims to learn the temporal consistency by learning alignment across various temporal granularities such as across different levels: video, clip, and frame. Furthermore, [39] utilizes the long-range frame-residuals along with the regular short clip to incorporate the long temporal context in contrastive learning. Besides that, Ranasinghe *et al.* [40] propose a dedicated SSL framework for video transformer where they create local and global spatiotemporal views of varying spatial sizes and frame rates from a given video. Its self-supervised objective aims to match features from these views of the same video, ensuring invariance to spatiotemporal variations in actions. Apart from that, the TeG [41] proposes learning across long-term and short-term temporal contexts by balancing the objectives through a weight coefficient. TATS [42] provides a complex solution where it tries to learn both consistency across the temporal context through maximizing mutual information between them and also discriminative features by identifying the playback rates which were used to achieve different temporal contexts.

Temporal Coherence

Temporal coherence in videos refers to the consistency and smooth flow of visual information over time within a video sequence. It means that successive frames in a video exhibit logical and continuous transitions, with objects and scenes changing in a manner that aligns with the laws of physics and real-world dynamics.

Some of videoSSL works which focuses on action specific downstream tasks try to learn temporal coherence. For example, in PRP [43], they utilize a dilated sampling strategy, enabling effective capture of temporal resolution. They reconstruct the original full-sequence from the dilated sampled sequence through a decoder, by doing so they claim to encourage the temporal-coherence in the learned representation. After that, TCE [44] encourages learning temporal-coherence in the contrastive learning framework, by taking the adjacent frame as the positive and frames from different video instances as the negatives. Although PRP and TCE show learning temporal coherence, they do not have huge success in showing high performance on semantic-level downstream tasks such as action recognition. Since temporal coherence deals with learning smoothness within the video rather than discriminating different videos, it does not help significantly in classifying the actions.

Recently temporal-coherence-based self-supervised objectives have seen more success in learning framewise video representations which are more useful in downstream tasks related to the intra-video temporal dynamics such as identifying the phases of an action (Details in Section 5.2). For instance, CARL [45] induces the temporal coherence between the frames of the videos, where it first passes the

 doi:10.20944/preprints202408.0133.v1

two overlapping clips of the same video to the video transformer network to get its framewise video representations. Now, to induce temporal coherence, the similarity between two frames of videos follows a smooth Gaussian prior which reduces the distance between frame indices. Similarly, in order to learn the temporally-coherent representations between the successive frames, VSP [46] puts the constraint that the framewise representation of the video should be modeled as a stochastic process. In this modeling, the action phase is considered as the goal-oriented stochastic process (Brownian bridge) and the framewise embedding from start to end is expected to follow a smooth transition.

While the above temporal-coherence-based SSL works do not require any labeled data, there are some works such as TCC [47], GTA [48] and LAV [49] also utilize self-supervised objective to learn the temporally-coherent representation for video alignment, however, they require video level action labels.

Temporal Correspondence

The notion of temporal correspondence — “what went where" [50] — is so fundamental for learning about objects in dynamic scenes, and how they inevitably change. Temporal correspondence deals with how the object/pixel/key points present in the current frame propagate to the other frames in the video. Since, the dense object/pixel level annotations are costly to obtain, learning temporal correspondence in self-supervised way is a very crucial problem.

One effective way to learn temporal correspondence is through cycle consistency. CRW [51] considers the video as a graph with image patches as nodes and affinities between them as the edges. In order to find the temporal correspondence between two points (nodes) in the video they optimize to get the strong edges from the random paths. Once the path is found, they encourage correspondence by learning to cycle back to the same source node from the target node. Extending the CRW for the multiple scale to enhance its fine-grained capability MCRW [52] introduces the hierarchy in the transition matrix computed between the pair of frames in a coase-to-fine-grained fashion.

Another well-known method, VFS [53] tries to learn temporal correspondence by learning the similarity between the two frames of a video. It forwards one pair or multiple pairs of frames from the same video into a Siamese network and computes the similarity between the frame-level features for learning the network representation. It does not use any negatives in the learning objective. Similarly, StT [54] also build upon the image self-supervised learning techniques where it proposes a spatial-then-temporal two-step training strategy. In the first step, it utilizes contrastive learning to initialize spatial features from the unlabeled images, whereas in the second step, it learns the temporal cues through the reconstructive learning.

3.1.2. Learning the Spatial Cues:

In this section, we discuss various methodologies employed to capture spatial information from videos. Spatial cues involve understanding the arrangement, position, and interaction of objects and elements within a two-dimensional space. This includes recognizing shapes, sizes, textures, and the relative positioning of objects within a single frame. Unlike temporal cues, which are concerned with how things change over time, spatial cues focus on the static aspects within each frame. The majority of these techniques in videoSSL draw inspiration from image-centric self-supervised learning approaches.

Spatial Order

Spatial order in videos refers to the structured arrangement and organization of visual elements, objects, or features within each frame or timestamp of a video sequence. It implies that the spatial configuration of objects and their relative positions in a frame carry important information about the actions that are being performed.

In the realm of image-based SSL, the approach of solving jigsaw puzzles emerged as one of the pioneering techniques, as demonstrated by Noroozi *et al.* [55]. Building upon this foundation, Ahsan *et*

 doi:10.20944/preprints202408.0133.v1

al. [56] propose solving a jigsaw puzzle as a pretext task to understand the spatial cues in videos. To achieve it, they split a frame into a 2×2 grid. For three video frames, it would be 12 patches per video, and those are randomly rearranged. During training, the network is taught to predict the order of these patches correctly. They use CaffeNet as their network and they also propose a unique approach to generate permutations efficiently. In another work [19], spatial permutation is also utilized just within a frame, where again the goal is to predict the correct spatial order of the patches within the frame from the set of disorganized patches.

Spatial Orientation

Spatial orientation refers to the arrangement, positioning, or alignment of objects or visual elements within a frame or scene in relation to a reference frame. It involves understanding the spatial relationships, angles, and orientations of objects with respect to each other or to a coordinate system.

Apart from solving jigsaw puzzles, rotation prediction is also a long-established approach in image-based self-supervised learning [57]. Inspired by this, [19,24,58,59] use spatial rotation for video representation learning. In this task, the video frame is rotated by four different orthogonal degrees: 0°, 90°, 180°, 270°. Then the model is asked to identify the degree of rotations in the input videos, which can be considered as a classification or regression task. In classification, the network predicts the rotation category; hence, rotations are predefined. In regression, the network predicts the rotation as a continuous variable. Furthermore, some works like Bai *et al.* [26], Zhang *et al.* [60] and Geng *et al.* [59] incorporate the technique of spatial rotation prediction as a supportive self-supervised objective to their overall learning objective.

Spatial Continuity

Spatial continuity, in the context of videos, refers to the property of maintaining a smooth and coherent visual flow or progression in the spatial domain. It involves the consistent arrangement and relationships of objects, features, or patterns within a video frame. MVD [61] introduces a novel approach based on masked image modelling. By manipulating spatial features and subsequently reconstructing them, their method effectively captures and learns meaningful data representations.

3.1.3. Learning the Spatiotemporal Cues

Although the previously discussed research has largely concentrated on either spatial or temporal cues in isolation within the realm of video self-supervised learning, this section delves deeper into pioneering studies that prioritize the integration of spatiotemporal cues. Recognizing that videos inherently possess both spatial patterns (the arrangement of objects and scenes) and temporal dynamics (the evolution of content over time), it becomes imperative to understand how these two aspects can be synergistically utilized for improved learning outcomes.

Spatiotemporal Order

Spatiotemporal order in videos refers to the organization and sequence of spatial and temporal information. Spatial information pertains to where things are in a frame, such as the position and orientation of objects or people, whereas temporal information, relates to the timing and sequence of events, capturing how objects or subjects move over time. When combined, spatiotemporal order helps in identifying and recognizing the content of video by understanding the movement patterns and interactions within a given environment. For instance, in a video of a person running, the spatial aspect would involve recognizing the person and their surroundings, while the temporal aspect would learn the movement of the person in successive frames.

Initial attempts in the spatio-temporal ordering based self-supervision is done by Buchler *et al.* [62]. In order to sample data permutations, which are fundamentally essential for any surrogate ordering task, a policy grounded in Reinforcement Learning (RL) is proposed. This policy requires a relatively smaller extra computational cost during the training process compared to the naive random

 doi:10.20944/preprints202408.0133.v1

perturbation. Consequently, this makes the process more efficient while maintaining the output robustness. Later another work from Kim *et al.* [63] proposing they a pretext task called 'Space-Time Cubic Puzzles' which requires a model to arrange permuted 3D spatiotemporal crops instead of 2D patches like VideoJigsaw [56]. Later, Zhang *et al.* [64] introduce the STTNet framework which incorporates a unique spatial self-supervised pretext task and a Transformer-based spatiotemporal aggregator to adaptively merge learned spatial and temporal features. Their network models spatiotemporal features by rearranging the sequence of frames temporally and altering the orientation of each frame spatially (for instance, by horizontally flipping them).

Spatiotemporal Continuity

Spatiotemporal continuity is defined as the consistency and progression of objects or features both in space (spatial) and over time (temporal). In videos, each frame is a spatial representation of objects at a given time. Spatiotemporal continuity involves understanding the movements of object through consecutive frames.

Most of the masking-based generative approaches rely on spatiotemporal continuity to learn representations. In masked autoencoder (MAE) based methods the main objective is to reconstruct either the patch (pixel-level) values or the latent space features, where the main constraint is learning the spatio-temporal continuity in the reconstructed output. One of the first works in the MAE for videos by Tong *et al.* [65] and Feichtenhofer *et al.* [66], they drop the random spato-temporal patches from the video and their main objective is to reconstruct the patches, which requires learning spatio-temporal continuity in the patch-level outputs. Building upon this, Wang *et al.* [67] propose a computationally efficient dual masking strategy. Furthermore, Yang *et al.* [68] emphasizes in learning motion by introducing reconstruction of the patches of frame-differences instead of simple RGB values, which requires learning the spatio-temporal continuity in the differences of the consecutive frames. Another work MAM^2 [69], combines both the objective of reconstruction of RGB patch and frame-difference patch in a disentangled fashion by utilizing two separate decoders. Developing in this area, [70] claims that prior MAE methods aim to predict the appearance of content in masked areas, but these approaches often fails to account for temporal aspects as content can often be inferred from a single frame. The work introduces Masked Motion Encoding (MME) which reconstructs both appearance and motion, addressing the two key challenges: representing long-term motion across multiple frames and obtaining fine-grained temporal clues from sparse video samples. Concurrently, Wang *et al.* [61] argue that existing MAE methods largely focus on reconstructing low-level features like raw pixel values and they proposed Masked Video Distillation (MVD) which is a two-stage framework that initially pretrains an image or video model by recovering low-level features of masked patches and then uses those features for masked feature modelling. Also, in order to leverage the strengths of different video teachers, a spatial-temporal co-teaching method is incorporated into the MVD, allowing distillation from both video and image teachers.

Apart from the above masking-based methods, [60] a pretext task based method also focuses on-learning the spatio-temporal continuity. It introduces a pretext task where the videos are transformed through spatiotemporal overlap-based data augmentation and task is to predict the spatiotemporal overlap rate (STOR), for which the model has to learn cues of spatiotemporal continuity in its representation.

3.1.4. Learning Low-Level Motion Cues

Optical flow determines a dense, pixel-wise correspondence between two consecutive images by identifying the second image's location of each pixel from the first image. This process generates a vector field that illustrates the apparent low-level motion or "flow" between the images. Optical flow estimation is a crucial challenge in computer vision, with improvements aiding various downstream applications including visual odometry, multi-view depth estimation, and video object tracking. The

 doi:10.20944/preprints202408.0133.v1

primary concept of self-supervised learning from low-level motion cues involves learning to predict optical flow from extensive unlabeled video datasets.

With this motivation SMURF [71], develops a student-teacher-based self-supervised framework, where the student model tries to predict the optical flow of the cropped frames which should give consistent output with the cropped optical flow of the pretrained RAFT [72] based teacher model. STAFNet [73] effectively decomposes video sequence motion into apparently matching regions and Low Matching Confidence (LMC) regions, utilizing a decoupled inference and training framework. The novel STAF block, a dynamic temporal model integrated with spatiotemporal context, adaptively judges and repairs LMC regions. Experimental results demonstrate its efficacy in significantly reducing endpoint errors.

3.1.5. Learning Geometric Cues

Geometry awareness in videos refers to the capability of algorithms to understand and interpret the geometric properties of objects and scenes in a video. This involves recognizing shapes, sizes, orientations, and the positional relationships of various elements within the frame. Geometry awareness is particularly crucial in tasks that require a deep understanding of the three-dimensional structure of the scene.

As one of the initial works in videoSSL, Gan *et al.* [74] proposed a fresh perspective of learning video representation by unlabeled geometry cues through a novel geometry-guided CNN. They utilize two types of free geometry data: optical flow from synthetic images and disparity maps from real 3D movies. These geometric cues effectively guide the CNNs to extract useful spatiotemporal features from the videos for the high-level semantic video understanding task. Apart from that, Sriram *et al.* [75] aim to learn the multi-view video representations and proposes a novel method called Homography-Equivariant Video Representation Learning (HomE) which explicitly models the representation space to maintain homography equivariance. Their main idea is to take a pair of frames from a multiple view of video and learn a representation that explains the tomographic relation between them. In a similar direction, Das and Ryoo [76] present viewCLR, where they learn the latent viewpoint representation through a view-point generator by optimizing a contrastive learning-based objective. They show that learning to generate complementary views of a video leads to useful video representation for the various viewpoints in the downstream tasks.

3.2. Learning Objectives with High-Level Cues

In this segment, we direct our attention towards methodologies that delve deeply into the acquisition of high-level spatiotemporal cues. Learning high-level cues in videos refers to the process of training algorithms to recognize and understand complex, abstract concepts and patterns within video data, beyond just the basic visual elements. High-level features are not directly observable like low-level features (such as skip rate, rotation, order, etc.), but rather they represent more sophisticated interpretations and inferences drawn from the raw data. This encompasses understanding and capturing semantic spatiotemporal information, distinguishing between different video instances, accounting for variations within instances, and discerning high-level similarities.

3.2.1. Discrimination Based Learning Objectives

Instance discrimination involves training a model to distinguish between different instances (e.g., video sequences/frames), aiming to distinctly separate them in feature space without using explicit labels. InfoNCE [77], a subset of instance discrimination, focuses on bringing similar (positive) pairs of instances closer in feature space while distancing dissimilar (negative) pairs. The key difference is that infoNCE specifically uses the strategy of comparing similar and dissimilar pairs to achieve instance discrimination, which can be attained through various other methods such as self-distillation approaches like BYOL [78] and SimSiam [79]. These techniques are crucial in video-related tasks like

 doi:10.20944/preprints202408.0133.v1

classification and activity recognition, as they enable the model to capture essential visual features and temporal dynamics without needing labeled data.

Much of the initial work in video self-supervised learning builds upon the simple image self-supervised contrastive learning methods. In the video domain, the InfoNCE-based self-supervised objectives were introduced through autoregressive generation modeling, which utilized contrastive learning to predict the next segments (frame/clip) of the video. Inspired by work from the image-domain Contrastive Predictive Coding [80], Han *et al.* [81] propose Dense Predictive Coding (DPC), which is enhanced by the use of a memory bank by the same researchers Han *et al.* [82] in MemDPC. Concurrent work [83] also utilizes a CPC-based objective including the hard-negatives from the same video instances.

Following the appearance of SimCLR [84], the video community adopted its success in utilizing the NT-Xent based implementation for videos. SimCLR considers two randomly augmented views of an image instance as positives and other instances as negatives. Similarly, in the video domain, two temporal crops (i.e., clips) from a video are considered as the positive pair, and clips from other videos as negative. Based on this observation, CVRL [85] builds positives by sampling two clips through their proposed sampling strategy. Feichtenhofer *et al.* [86] also show an improved variant with SimCLR where they consider multiple positive clips from a video instance instead of just two clips. Another successful approach adopted from the image SSL domain is MoCo [87], which does not require the large batch size of negatives like SimCLR. VideoMoCo [88] extends the MoCo framework to video by incorporating temporal dynamics into contrastive learning. It utilizes adversarial learning for temporal augmentation of video sequences and employs a temporal decay mechanism to progressively decrease the influence of older keys in the queue, enhancing the temporal relevance of the learned representations.

Another predominant image SSL approach—BYOL (Bootstrap Your Own Latent)—differentiates itself from SimCLR and MoCo by eliminating the need for negative pairs in training. This model relies on a unique dual-network architecture, where a "student" network learns to predict the output of a simultaneously updated "teacher" network, effectively using the teacher's output as a dynamic target. This method allows BYOL to avoid the complexities and potential biases associated with negative sample selection, which can be especially beneficial in managing the diverse and nuanced temporal patterns found in video data. Feichtenhofer *et al.* [86] successfully extend BYOL to videos and call it ρBYOL, where they maximize mutual information between distant clips from the same video. They also successfully show results with SWaV [89], MoCo, and SimCLR, however, observing that frameworks like BYOL and SwAV, which do not require negative samples, offer significant advantages in simplicity and effectiveness, particularly by avoiding the complexities associated with negative sample selection in the learning of temporally persistent features in videos.

Another approach in image SSL is Deep InfoMax (DIM) [90] and its extension Augmented Multi-scale Deep InfoMax (AMDIM) [91], which are self-supervised learning approaches that fundamentally differ from SimCLR by focusing on maximizing mutual information between different parts of a single input (rather than contrasting augmented views of an input as in SimCLR). Specifically, DIM aims to maximize the mutual information between global features and local features within an image, encouraging the model to learn comprehensive representations that encapsulate both high-level and detailed aspects of the visual data. AMDIM enhances this by applying the concept across multiple scales and augmentations, thus further enriching the feature learning through diversity in representation at different resolutions and transformations. The approach has been adapted to the video domain by utilizing local views derived from spatio-temporal features [18,92]. Implementations of various extensions of the image SSL methods are provided by Feichtenhofer *et al.* [86] [1] and Sarkar *et al.* [93] [2].

[1] https://github.com/facebookresearch/SlowFast/tree/main/projects/contrastive_ssl
[2] https://github.com/pritamqu/OOD-VSSL/tree/master/codes/vssl-train

 doi:10.20944/preprints202408.0133.v1

Since the image-based contrastive and self-distillation objectives do not emphasize temporally focused learning, some methods utilize these SSL objectives with temporal pretext tasks such as playback rate prediction in PacePred [30], TaCo, temporal order prediction [20,21,33], etc. Jenni and Jin [28] introduce a temporal-equivariance property in addition to the regular contrastive loss. They first classify the temporal relation between two clips from the video into overlapping, ordered, and unordered classes, concatenate the features of two clips to get a video-level feature, and apply the regular instance-level contrastive loss.

In learning temporal features through handcrafted temporal pretext tasks, some methods explicitly focus on learning temporal features through contrastive learning or high-level cue-based SSL objectives. One of the pioneering works in this domain is the Temporal Contrastive Learning (TCLR) framework [94]. It formulates the contrastive loss using negatives from temporally non-overlapping clips of the same video. This marks the first attempt to explicitly learn temporal distinctiveness using a higher-level objective like contrastive loss. They propose two temporal contrastive losses to encourage temporal distinctiveness at two different levels of temporal aggregation: (1) clip-level and (2) feature-level pooling layer. Similarly, Wang *et al.* [95] proposed the Temporal Discriminative Learning (VTDL) framework, which employs triplet loss to discriminate between temporally misaligned clips. Chen *et al.* [96] introduce a unique approach known as intra-video contrastive learning (intra-VCL). It utilizes an asynchronous long-term memory bank that stores representations of all video snippets, thereby facilitating the discovery of additional positive or negative snippets within a video to enhance the contrastive learning process. Similarly, inter and intra-video instance-based contrastive losses are also utilized by [97,98].

3.2.2. Similarity Based Learning Objectives

Clustering, as applied in machine learning, groups objects based on inherent similarities, differing significantly from instance discrimination-based objectives typically used in self-supervised learning. While instance discrimination trains models to distinguish and separate individual instances in the feature space, clustering seeks to group similar instances together, enhancing intra-class compactness and inter-class separability. This makes clustering particularly suited for tasks where the preservation and understanding of group characteristics are crucial. In the context of video, clustering leverages both spatial and temporal data to detect and categorize content that shares common features, thus providing a more holistic approach to video analysis compared to instance discrimination which focuses on identifying distinct, isolated features.

Building on the concept of clustering for video representation, Zhu *et al.* in [99] highlight the efficacy of clustering in grouping similar videos, which improves video representations through enhanced intra-class compactness. They emphasize the iterative nature of clustering and representation learning, advocating for a continuous refinement process until convergence. In their proposed two-stage framework, Instance-CL and unsupervised clustering combine to progressively refine temporal representations, addressing the challenges posed by initial clustering outcomes. Further advancing this field, Khorasgani *et al.* [100] introduce Self-Supervised Learning with Iterative Clustering (SLIC), employing iterative clustering to group video instances and using pseudo-labels for sampling challenges in positive and negative instances. This method marks a pioneering approach in robust representation learning via iterative clustering. Additionally, Miech *et al.* [101] explore semantic analysis by aiming to develop a joint embedding space that correlates the semantic congruity between text descriptions and video content, transforming raw video pixels and text into coherent representations for deeper comparative analysis. Also employing clustering, Tokmakov *et al.* [102] apply this method around motion trajectories to enhance video analysis through Improved Dense Trajectories and 3D Convolutional Neural Networks.

 doi:10.20944/preprints202408.0133.v1

3.3. Learning with Precomputed Visual Priors

Visual priors refer to precomputed input modality that capture essential visual information from video data. These priors provide a rich source of information for understanding the content and dynamics of videos, making them valuable for self-supervised learning in video understanding. By leveraging these precomputed visual priors along with the regular RGB stream, models can learn to recognize patterns, movements, and relationships within the video without the need for explicit annotations. Examples of commonly used visual priors in video self-supervised learning include optical flow, frame residuals, and improved dense trajectories (IDT). Each of these priors captures different aspects of video data:

3.3.1. Optical Flow

Optical flow is a representation of the apparent motion of objects between consecutive frames in a video. It is computed by estimating the displacement of pixels between frames, which provides information about the direction and speed of moving objects. One popular method for computing optical flow is the TV-L1 [103] algorithm, which is known for its robustness and accuracy in capturing fine-grained motion details. Another faster way to estimate optical flow using the neural network is through RAFT [72].

Wang *et al.* [104,105] learn statistics from two streams: appearance (RGB) and motion (optical flow). They initially partition multiple frames into grids, identifying the most motion-intensive grids using the magnitude of optical flow and the most color-rich grids using appearance information. The distribution of such patches provides the self-supervisory signal in their approach. Subsequently, Xiao *et al.* [106] propose MaCLR, a two-stream (RGB, flow) contrastive learning framework that trains with three simple InfoNCE objectives similar to SimCLR: (1) InfoNCE within the RGB modality alone, (2) InfoNCE within the Flow modality alone, and (3) cross-modal InfoNCE where positives are formed using clips from the same video of RGB and flow, and negatives are formed using clips from different videos. Building on this, Ni *et al.* [107] introduce MSCL framework, which utilizes RAFT [72] to extract flow and samples video regions with the highest motion. This method distills the knowledge from flow to RGB through a frame-level contrastive objective, complemented by regular clip-level contrastive learning. Additionally, GOCA [108] introduces a dual-view clustering strategy that uses initial cluster assignments from one view to guide the clustering in the other view, effectively synchronizing the cluster structures across RGB and optical flow modalities. This strategy not only enhances the semantic coherence of clusters but also robustly counters the inherent noise in individual views. Furthermore, it introduces a unique regularization strategy designed to prevent feature collapse, a common issue in cluster-based self-supervised learning frameworks. Lastly, the CoCLR method by Han *et al.* [109] introduces a novel self-supervised co-training strategy that utilizes the complementary nature of RGB and optical flow views to enhance the self-supervised objective. This method employs Multi-Instance InfoNCE loss to mine informative positive pairs across different views, effectively refining video feature representations through iterative cross-modal feedback, thereby demonstrating improved performance.

3.3.2. Motion Trajectories

Tokamov *et al.* [102] present a heuristic-based Improved Dense Trajectories (IDT) descriptor approach which clusters motion trajectories. Improved Dense Trajectories (IDT) refine the dense trajectory approach by filtering out irrelevant motion paths and incorporating additional descriptors such as Histograms of Oriented Gradients (HOG), Histograms of Optical Flow (HOF), and Motion Boundary Histograms (MBH). This enhances the capture of detailed motion and appearance data across video frames. Tokmakov *et al.* [102] leverage Improved Dense Trajectories (IDT) to infuse temporal knowledge into the conventional instance-level contrastive objective, which typically lacks this aspect. Furthermore, compared to existing masking-based methods, [70] proposed a novel Masked Motion Encoding method (MME), which uses reconstructing motion trajectory to learn high representations.

 doi:10.20944/preprints202408.0133.v1

They utilize pre-extracted optical flow-based tracking points to compute trajectories, which are then employed in the decoder for reconstruction purposes.

3.3.3. Silhouettes

A silhouette in video processing refers to the outline or general shape of an object, typically a person, extracted from the background to emphasize the movement and posture without including appearance details. This focus on silhouettes is crucial for gait recognition as it allows the model to learn and identify individuals based on their movement patterns rather than their appearance, which is vital for maintaining effectiveness regardless of clothing or lighting conditions.

In this line, GaitSSB [110] is a specialized videoSSL method designed to learn gait recognition from unlabeled walking videos. In the frame-level preprocessing, silhouettes of pedestrians are extracted using a segmentation model [111], followed by morphological operations. For the SSL pretraining, the method focuses solely on silhouette sequences and applies a regular InfoNCE loss, where positives are formed from two clips of the same video and negatives from clips of different videos. It also proposes silhouette augmentation operator (SAO) to apply to the clips before infoNCE loss.

4. Challenges, Issues and Proposed Solutions in Video SSL

4.1. Data Utilization Strategies

In the realm of video-based self-supervised learning (SSL), the quality and quantity of data play pivotal roles in determining the effectiveness of the learning process. Videos, by their nature, are data-intensive, often requiring substantial storage and computational resources. In this section, we delve into two critical aspects of data management in video-based SSL: addressing the issue of data scarcity and enhancing the quality of data samples

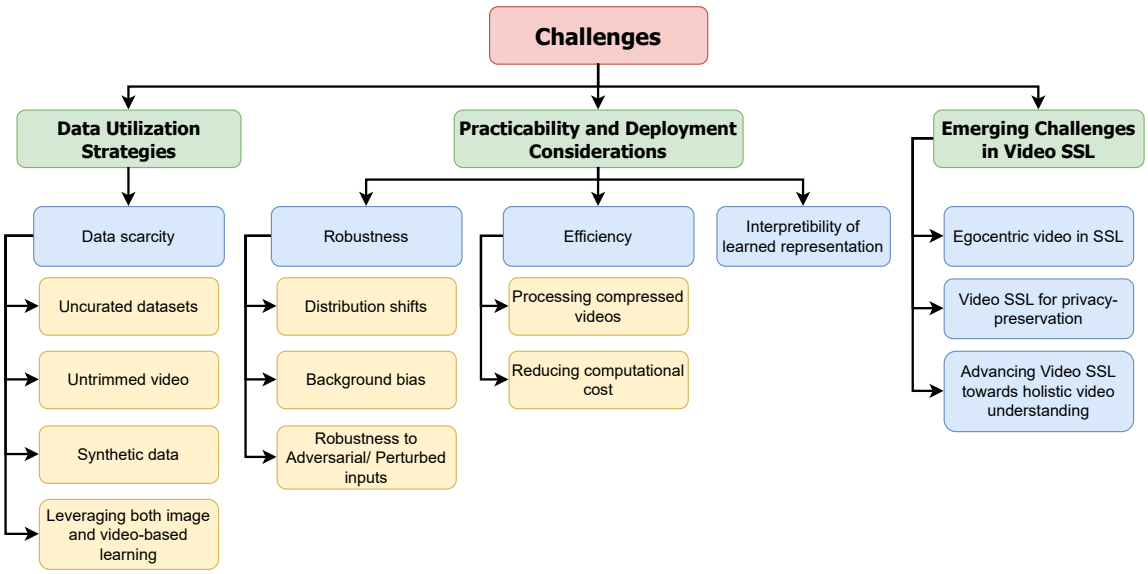


Figure 4. Challenges and Issues in Video SSL (Details in Section 4)

4.1.1. Data Scarcity

SSL presents a viable pathway for overcoming the limitations imposed by the need for extensive labelled datasets by leveraging unlabelled data to learn the underlying representations of videos. Nonetheless, the efficacy of SSL is contingent upon the availability of substantial datasets to facilitate the learning process. The acquisition of large datasets, even in the absence of labelling, can be a costly process. In this section, we discuss three main approaches which address the data scarcity issue

 doi:10.20944/preprints202408.0133.v1

mainly: learning from uncurated datasets, leveraging external data, learning from prior knowledge and utilizing synthetic data.

Learning from Uncurated Datasets

Most video self-supervised learning methods utilize unlabeled versions of curated datasets such as Kinetics and UCF101. These datasets are prepared from original internet videos by carefully extracting moments where the actions take place. However, to fully leverage the large corpse of unlabeled data such as internet videos, video SSL methods need to be capable of learning from uncurated videos.

In the early stages, XDC [112] demonstrates competitive performance when applied to uncurated datasets. Although the authors did not primarily focus on uncurated data, their work laid a foundational groundwork in this area in video self-supervised learning. Buidling on this foundation, [113] introduces an innovative method for leveraging uncurated datasets by examining the efficacy of representation learning from movies and TV shows, demonstrating the significant effectiveness of such learning from these uncurated sources. Following up [114] continued this exploration, with a focus solely on movies. They posited that, despite the high level of production and artistic curation in movies, they remain semantically uncurated. Identifying unique attributes inherent to movies, they thoroughly investigated how employing negative sampling within the content itself could enhance the performance of contrastive learning methods.

Using Untrimmed Videos

Current video datasets consist of trimmed video data and these trimmed videos are obtained from manually labeled untrimmed videos. Real untrimmed videos consist of both background and foreground frames. Hence, [23] proposes a method called 'exploring relations in untrimmed videos' (ERUV), which learns features involves creating relationships between video clips, which allows combining existing self-supervised techniques with their custom-designed connections. Taking a different approach, HiCo [115] authors point out that there are two main issues using trimmed videos: the limited diversity in visual patterns and performance gain due to the reliance on manually trimmed videos. Hence, HiCo leverages more abundant information in untrimmed videos to learn a hierarchy of consistencies in videos, such as visual consistency and topical consistency. Improved upon this [116], proposes the HiCo++ framework, which represents a significant evolution from its predecessor, HiCo, by offering greater flexibility in sampling multiple visually consistent pairs from each untrimmed video. This capability marks a notable improvement over the original HiCo framework, which was limited in its sampling capabilities. Furthermore, the authors have enhanced the evaluation of topic consistency. They achieve this by aggregating visually consistent features, which contributes to the training stability of topical consistency learning.

Using Synthetic Data

Synthetic data is artificially generated data, as opposed to data collected from real-world events. One of the main advantages of synthetic data is that it can be generated to include a wide range of scenarios and conditions that might not be easily obtained from real-world data, allowing for more comprehensive training and testing of models.

Li *et al.* [117] deal with the dense temporal correspondence tasks, where they utilize the supervised training from the synthetic data since it is easy to obtain label on that. For the unlabeled data, they utilize generic reconstruction loss and to bridge the domain gap between the real and synthetic domain, they utilize an adversarial loss for the domain invariant representation. This approach marks a significant advancement in the utilization of synthetic media for detailed and nuanced feature extraction, showcasing a novel methodology that combines the strengths of both artificial and authentic visual sources for improved learning outcomes. Also, [118] contributes to this evolving field by suggesting the use of synthetic motion trajectories, or 'tubelets,' to delve into motion-centric

 doi:10.20944/preprints202408.0133.v1

video representations. This approach represents a shift towards understanding video content through synthesized motion patterns, adding a new perspective to the study of video SSL.

Leveraging Both Image and Video-Based Learning

One effective way to utilize available video data is to understand it in terms of both the video and image domains. Motivated by this, many videoSSL methods employ two streams of inputs from the video, utilizing both image-based and video-based SSL learning to enhance video understanding.

Kong *et al.* [119] believe that in video representation learning, both the holistic video and its constituent frames play pivotal roles. To achieve this they propose a cycle contrastive learning loss method that encourages these characteristics in video representation. Their motivation is that within the domains of videos and their constituent frames, the representations should exhibit a closeness between them. Concurrently, these representations should maintain distinctiveness from all other videos and frames within their respective domains. Building on this concept, [120] proposes a novel video / image for visual contrastive learning of representation framework dubbed Vi^{2}CLR. As the name suggests, this SSL framework is able to concurrently interpret both image and video (2D and 3D) representations, capitalizing on both the dynamic and static visual indicators and instances of similarity and dissimilarity. Their newly designed neural network architecture introduces two separate convolutional neural networks tailored to adeptly handle visual recognition challenges across both video and image domains. In a similar vein, [61] proposes a technique involving masked video distillation that is further enhanced with an effective co-teaching strategy. This approach derives advantages from the joint contributions of both images and videos.

4.2. Practicability and Deployment Considerations

Studying the resilience of learned SSL video representations is crucial before deployment. We categorize developments that address four main challenges: distribution shifts, background bias, and adversarial resilience.

4.2.1. Robustness

To study robustness related to the VideoSSL, we categorize the necessary developments to tackle four main challenges: distribution shifts, background bias, feature decomposition, and the enhancement of adversarial resilience and learning.

Distribution Shifts

When addressing distribution shifts, the focus is on ensuring model performance remains stable and reliable when faced with data that may differ from the training distribution. In [121], the proposed approach ViTTA is tailored to spatiotemporal models and consists of a feature distribution alignment technique that aligns online estimates of test set statistics towards training statistics. Compared to the ViTTA framework, Sarkar *et al.* [93] investigate the behavior of video SSL methods under different forms of distribution shift that commonly occur in videos due to changes in context, viewpoint, actor, and source. The authors introduce a comprehensive out-of-distribution (OoD) test bed curated from existing literature to evaluate the robustness of video models.

Background Bias

Mitigating background bias pertains to the ability of model to distinguish and maintain focus on the primary objects or features of interest, despite changes or variations in the background. This is crucial for video SSL models, where the background can often be dynamic and unpredictable. In the literature, there are two approaches to avoid scene/background bias:

The first approach smartly utilizes augmentations such as mixmatch, cutmix, etc., in a spatio-temporal manner to reduce the background bias. Examples include DSM [122], which uses Spatial Local Disturbance and Temporal Local Disturbance to enhance model focus on motion by creating and differentially aligning positive and negative video clips in latent space to the original clip. Background

 doi:10.20944/preprints202408.0133.v1

Erasing (BE) method introduced in [123] eliminates irrelevant features or 'background noise' from the training data. Other studies in this approach include [124–130].

The second set of methods utilizes effective motion cues to reduce scene bias. For instance, [131] leveraged the P-frames of compressed videos to mitigate the scene bias. Additional contributions in this area include [126,132–134].

Robustness to Adversarial/Perturbed Inputs

In the context of video self-supervised learning, robustness to adversarial and perturbed inputs is a critical aspect to ensure the reliability and security of the learned representations. Adversarial inputs are deliberately crafted perturbations designed to deceive the model, while perturbed inputs refer to inputs with added noise or distortions that occur naturally or are artificially introduced. Both types of disturbances can significantly impact the performance of video SSL methods, especially in downstream applications where the integrity of the input is crucial for accurate decision-making.

Adversarial Attacks: Adversarial attacks involve generating inputs that are specifically designed to cause the model to make errors, while perturbations can be random or structured noise added to the video data. Contrastive self-supervised learning (CSL) has demonstrated the ability to equal or exceed the results of supervised learning in categorizing images and videos. Yet, there is still a significant lack of understanding as to whether the two learning methods produce similar types of representations. [135] explores this issue by examining adversarial resilience. Their investigation has discovered that CSL inherently has a greater vulnerability to disruptions than supervised learning and pinpointed the even dispersal of data representation across a unit hypersphere within the CSL representational space as the primary factor leading to this occurrence. Their work presents the first comprehensive evidence of the increased susceptibility of CSL to changes in input, substantiated by rigorous tests for image and video categorization. Compared to the above research in adversarial resilience [88] focuses on adversarial learning to improve the temporal robustness of the encoder. Their architecture focuses on using MoCo frame work for videos and the temporally adversarial learning and the temporal decay improve the temporal feature representation of MoCo.

Input Perturbations: Input perturbation refers to the corruption of the input video by real-life-like noises such as Gaussian, Shot, Impulse, and Speckle noise. Dave *et al.* [33] investigate the performance of video retrieval methods under temporally-inconsistent noises. They demonstrate that their method achieves superior performance compared to other video SSL methods under such input perturbations, owing to the frame-independent spatial jittering augmentation employed in their self-supervised objective.

4.2.2. Efficiency:

Efficiently processing video data remains a crucial objective in reducing operational costs, given the substantial size of video files. Video data, due to its high dimensionality and the vast amount of information it encapsulates, necessitates significant computational resources for processing and analysis. Thus, optimizing the techniques and algorithms for handling video data is imperative to minimize the required storage space and computational power, ultimately leading to a reduction in associated expenses.

Processing Compressed Video

Among the different challenges associated with handling and processing videos, an important one is the significant computing and storage requirements of the conventional methods of video processing that require frames to be decoded before being processed. The work [136] proposes to eliminate the costly decoding step. They use compressed video format to learn video representations directly, leveraging the inherent atomic nature of compressed videos in Group of Pictures (GOP) structures and their multi-modal representation capacity. To further strengthen their approach, they propose two specific pretext tasks: predicting motion statistics within a spatiotemporal grid structure and

 doi:10.20944/preprints202408.0133.v1

predicting correspondence types between I-frames and P-frames following temporal transformation. In [131], they take advantage of compressed videos to decouple the motion and context information and both of these information can be efficiently extracted at over 500 fps on CPU. They are the first ones to introduce a method that utilizes the different modes in compressed videos as an effective source of supervision for learning visual representations.

Reducing Computational Cost

As natural human action videos exhibit gradual changes from frame to frame, they typically possess a high degree of spatio-temporal redundancy, allowing for many such patches to be omitted without significant loss of information. Recent MAE-based videoSSL methods [65,66] exploit this by dropping around 90% of the patches, achieving high performance in finetuning. However, this approach presents several challenges: (1) it demands substantial computational resources due to a lengthy decoder; (2) random patch dropping necessitates extended pretraining periods; (3) it struggles in downstream tasks without comprehensive finetuning, which is computationally intensive as it requires processing all patches.

Several recent efforts have addressed these issues. VideoMAE-V2 [67] introduces a dual masking strategy that reduces the memory demands of the decoder. Instead of randomly dropping patches, newer methods opt for selective patch dropping to further reduce memory requirements. For instance, [137] selects tokens linked to moving objects, [138] bases patch selection on optical flow priors, and [139] introduces Cell Running Masks (CRM) that maintain spatio-temporal correlations while dropping patches. To circumvent the need for full finetuning during downstream tasks, V-JEPA [12] employs a reconstruction strategy in the latent space which is also learned through an additional encoder (obtained via a moving average of the model). It demonstrates effective downstream performance on frozen model by simply adding a learned attention layer.

4.2.3. Interpretibility of Learned Representation:

Interpretability in videoSSL involves the ability to describe the internal mechanisms and learned features of SSL models. This aspect is essential as it helps determine which downstream tasks a videoSSL model is best suited for, ensuring that the model captures meaningful patterns relevant to specific applications rather than exploiting arbitrary dataset characteristics. While we discuss most details in the experimental section, here we provide a higher-level overview. VideoSSL methods that focus on learning temporal correspondence [53] are particularly well-suited for tasks such as video object segmentation and pose tracking.

SSL objectives based on higher-level cues, such as contrastive learning, are more complex to interpret. Recently, TimeBalance [140] has categorized contrastive learning-based videoSSL objectives into two types based on their representations: Temporally-Invariant and Temporal-Distinctive. Predominantly, contrastive methods [85,86] consider clips from the same video as positives and repel clips from different videos, promoting temporally-invariant representation learning. In contrast, methods like [94,97,98] use temporally non-overlapping clips from the same video as negatives, promoting temporally-distinctive representations. It remains an open question whether temporally invariant or distinctive representations are more effective. [140] presents a comparison of both types in action recognition tasks. As shown in Fig. X, temporally invariant methods are more effective in repetitive actions or scenes with a dominant setting, such as fencing, whereas temporally distinctive methods excel in complex actions involving multiple components, like the javelin throw. While [140] offers a solution for semi-supervised learning setups, developing a self-supervised learning method that can effectively learn both types of representations within a single model remains a significant research challenge. Additionally, interpreting methods based on masking i.e. MAE, remains an unexplored area of research.

 doi:10.20944/preprints202408.0133.v1

4.3. Navigating Emerging Challenges in Video SSL

In this segment, we delve into a selection of emerging challenges in video-based self-supervised learning that have garnered substantial attention from the research community. We aim to provide a comprehensive overview, highlighting the primary areas of focus and the nuances associated with these emerging issues. By doing so, we intend to shed light on the current landscape of research and underscore the significance of addressing these challenges to foster innovation and progress in various fields of study.

4.3.1. Egocentric Video in SSL:

Videos from an egocentric, or first-person, view provide a unique perspective that mimics an individual's point of view. Analyzing these videos can be challenging due to constant movement and shifting focal points, making research in this area particularly complex.

The development of videoSSL methods for egocentric content is still in its early stages. Escorcia *et al.* [141] introduce an Objects In Contact (OIC) representation, utilizing video object regions detected by an off-the-shelf hand-object contact detector, which demonstrates high performance across multiple egocentric video classification datasets. Another approach by Xue *et al.* [142] learns the temporal alignment between the object-encoding of the third-person and egocentric views using a dynamic time warping (DTW)-based objective, achieving robust performance on downstream tasks related to action phases. From both studies [141,142], it appears that object-centric representations are a promising direction for advancing egocentric videoSSL methodologies.

4.3.2. Video SSL for Privacy-Preservation:

Privacy-preserving video understanding aims to mitigate the leakage of private information while maintaining the utility of video downstream performance. SPAct [143] proposes a self-supervised method, which removes all spatial cues in videos by minimizing the similarity of frames within the same video. Simultaneously, useful utility features are preserved through an action recognition utility branch. Their approach achieves performance comparable to supervised adversarial learning-based privacy-preserving approaches. For extending SPAct to the anomaly detection downstream task, Fioresi *et al.* [144] propose Ted-SPAD, which adapts the self-supervised privacy removal objective of SPAct to longer videos by introducing temporal distinctiveness alongside proxy anonymization task.

While the aforementioned methods preserve privacy by learning an anonymization function, another approach, Rehman *et al.* [145], employs a Federated Learning-based strategy to enhance the safety of video self-supervised learning and prevent privacy leakage by promoting a decentralized federated learning paradigm. To achieve this, they propose FedVSSL, which integrates different aggregation strategies and partial weight updating.

4.3.3. Advancing Video SSL Towards Holistic Video Understanding:

The human visual system is highly sophisticated, capable of understanding complex scenes, recognizing objects, perceiving depth, and interpreting motion with remarkable efficiency and accuracy. Current videoSSL methods have focused primarily on one family of downstream tasks, either high-level semantic recognition or low-level temporal correspondence. This specialization indicates a significant limitation in the scope of videoSSL methods. However, recent efforts have aimed to address these limitations.

NMS [33] effectively achieves low-level temporal correspondence by disrupting the shortcuts in temporal pretext tasks, while also enabling semantic understanding through a video-level objective. V-JEPA [12] demonstrates capabilities in semantic downstream tasks related to videos and images without requiring changes to the architecture of the learned model. MC-JEPA [146] employs both a appearance based distillation objective and a low-level optical-flow reconstruction as SSL objective in a multi-task setting, performing adequately on both appearance-based semantic video tasks and motion-related optical flow estimation tasks. Despite these attempts, none of the methods perform

 doi:10.20944/preprints202408.0133.v1

on par with expert videoSSL methods for each family of downstream tasks, highlighting the ongoing challenge of achieving human-like unsupervised video intelligence.

Furthermore, earlier attempts to mimic human cognitive understanding in sources like [147–150] provide valuable insights into their SSL objectives. However, they fall short in performance on regular downstream tasks, which reveals a significant research gap in their SSL objectives' ability to mimic the human cognitive ability to process video.

5. Experimental Comparison

In our experimental evaluation, we delve into various evaluation protocols applied to a range of downstream tasks. Distinguishing our work from prior surveys, such as Schiappa *et al.* [11], which focused solely on action-related semantic downstream tasks, we present a comprehensive study encompassing a broad spectrum of tasks from high-level semantic recognition to low-level temporal correspondence as shown in Figure 1. To our knowledge, this represents the first extensive exploration of videoSSL methods across such a diverse array of downstream task families, spanning the complete spectrum from high to low-level video processing tasks. For semantic recognition tasks, our focus includes not only standard action recognition but also multi-label video classification of the video attributes in Section 5.1. We then shift our attention to intra-video semantics in Section 5.2, such as action phase recognition, crucial for the class-agnostic fine-grained understanding of an action. On the low-level understanding side in Section 5.3, our study examines temporal correspondence at various levels including object, keypoint, and pixel levels, such as in optical flow estimation. Lastly in Section 5.5, we assess the robustness of these videoSSL methods against distribution shifts and input perturbations.

5.1. Semantic Understanding Based Tasks

Understanding video content at this level enables systems to interpret complex scenarios and make informed decisions based on both the context and dynamic nature of the content.

5.1.1. Action Recognition

Action-related downstream tasks require a spatio-temporal understanding of the video content to identify human actions. In the realm of action recognition, we investigate the task under three different parameter-tuning protocols: Linear Evaluation, Full Fine-tuning, and Zero-Shot Action Retrieval. Each setting offers unique insights into how well videoSSL methods can adapt to new datasets and generalize across different levels of supervision. We cover the most commonly used datasets covering various scenarios such as UCF101 [151], HMDB51 [152], and Kinetics400 [153] , which are composed of internet videos showing mostly third-person actions. Something-SomethingV2 [154] covers action categories with various hand-object interactions, whereas, NTU60 [155] is collected from controlled scenarios with different actions occurring against the same background.

Most commonly, videoSSL methods utilize the unlabeled version of Kinetics400 as the pretraining dataset and perform downstream tasks under various protocols. Firstly, we examine the zero-shot video retrieval protocol, where the goal is to retrieve the video from the search space that has the same action class as the given query video without any finetuning. The most commonly used downstream datasets for this task are UCF101 and HMDB51, where their test sets are used as the query videos and training sets are utilized as the gallery videos. The results for video retrieval are shown in Figure 5. Secondly, we present the results on action recognition by training a linear layer on top of the frozen SSL backbone, as depicted in Figure 6. Finally, by finetuning the entire model, we present full-finetuning results in Figure 7.

 doi:10.20944/preprints202408.0133.v1

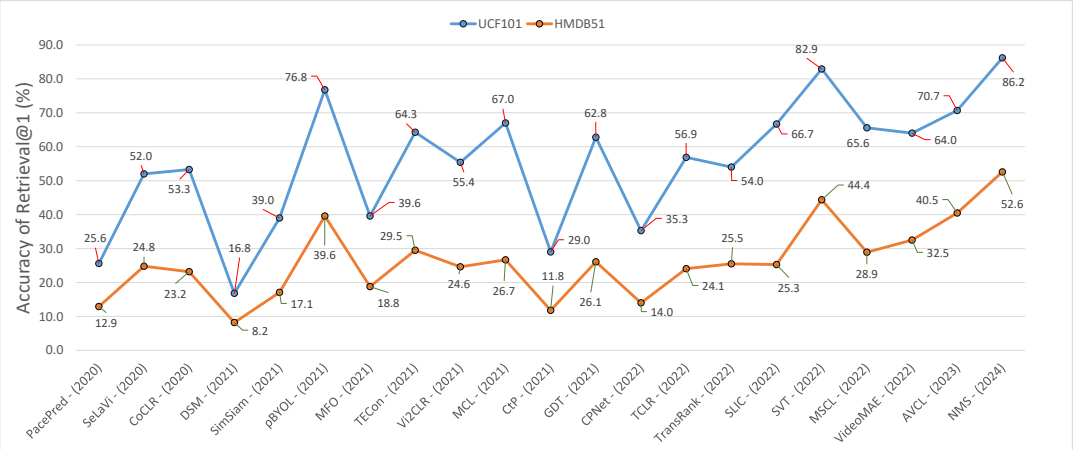
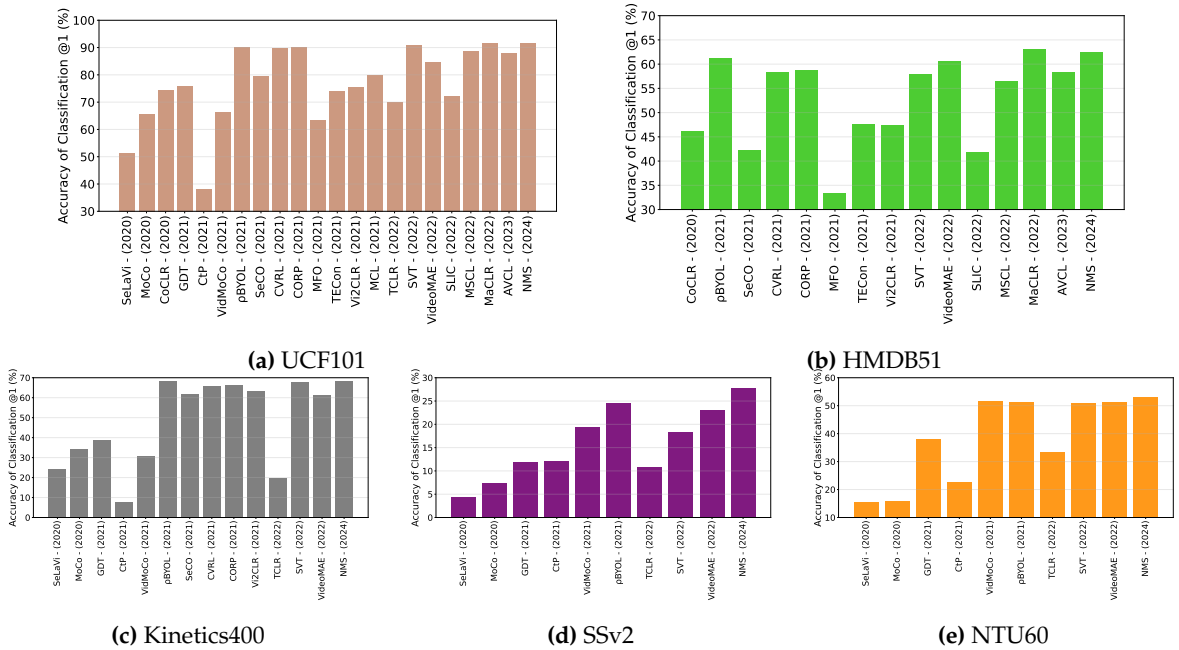


Figure 5. Zero-shot Video-to-Video Retrieval Performance of VideoSSL methods (in chronological order) on UCF101 and HMDB51



(a) UCF101 **(b)** HMDB51

(c) Kinetics400 **(d)** SSv2 **(e)** NTU60

Figure 6. Linear Action Classification Performance of VideoSSL methods (in chronological order) on UCF101 and HMDB51

Our initial observations reveal that videoSSL methods performing well in zero-shot action retrieval (Figure 5) tend to exhibit comparable performance in linear evaluation (Figure 6), as seen with recent methods such as NMS [33] and SVT [40]. However, some methods like VideoMAE [65] excel only in full-finetuning, indicating that the video representations learned through masked image modeling objectives are not directly suitable for downstream tasks without comprehensive finetuning. Additionally, a close examination of the finetuning code in publicly available repositories reveals that some methods [65,86] incorporate non-standard components like label smoothing and exponentially moving average weights to enhance finetuning performance, which compromises fairness in comparisons with the standard finetuning protocol [82] followed by most methods. Considering the expectation for videoSSL models to assist in video search applications, zero-shot action retrieval on the frozen pretrained model is more appropriate.

Furthermore, most datasets used for evaluating action recognition are coarse-grained and do not require detailed spatio-temporal understanding, for example, datasets like FineGym [156] and Diving48 [157]. Recently, SEVERE [158] introduced a comprehensive benchmark for action-related

 doi:10.20944/preprints202408.0133.v1

downstream tasks, considering various factors such as domain and number of samples, which could be crucial for demonstrating the efficacy of a videoSSL method in action-related downstream tasks.

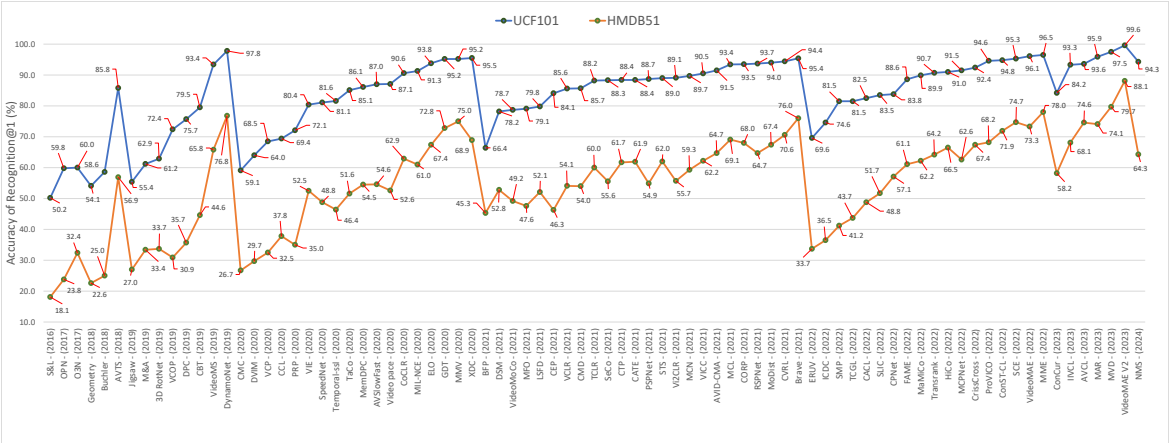


Figure 7. Full-finetuning Performance of VideoSSL methods (in chronological order) for action recognition on UCF101 and HMDB51

5.1.2. Video Classification

Video classification extends beyond action recognition to encompass a comprehensive analysis of every distinguishable element within the video frame, including objects, scenes, and events. This task is pivotal in applications where a holistic understanding of the video content is necessary, such as in content recommendation systems, event detection in surveillance, and contextual advertising. In order to perform the reliable evaluation, large-scale Holistic Video Understanding (HVU) dataset [159] is utilized, which provides multi-label annotations for each aspects of the video: scenes, objects, actions, attributes, and concepts. In this protocol, different linear classifiers are trained over the frozen features, and performance is reported in mean average precision (mAP).

From the results shown in Figure 8, it appears that methods which utilize the imagenet self-supervised pretrained weights such as in [33,40] obtain better results than other methods in all video semantics. The main limitation of the current multi-label video classification studies only consider one large-scale dataset HVU and only in linear evaluation setting, however, in the future, it should be studied on multiple datasets such as Coin [160] and APPROVE [161] in different setting such as zero-shot and full-finetuning as well.

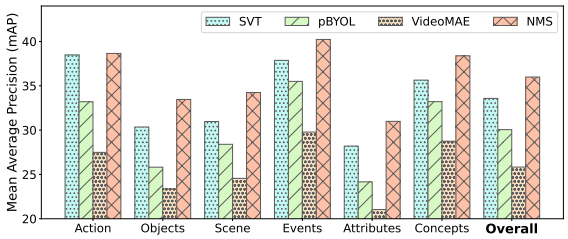


Figure 8. Video Attribute Recognition on multi-label video classification benchmark of HVU dataset.

5.2. Action Phases Related Downstream Tasks

This series of downstream tasks mainly focuses on the temporal dynamics within one video such as action phases rather than identifying action at video level. Action phase can be defined as integral part of an action and action phases follow a deterministic order to complete an action. For example, in the action of `Baseball Pitching`, it follows a deterministic order of action phases such as taking a stance, stretching the arm, and throwing. The most common used datasets which have the action phase annotations are PennAction [162] and Pouring [163] There are three most common tasks to evaluate framewise video SSL representation for the action phases:

 doi:10.20944/preprints202408.0133.v1

5.2.1. Phase Classification

After the SSL pretraining, the backbone is frozen and a linear classifier is trained on the action phases for a specific action (e.g., `Baseball pitching`). For the test set, each frame is classified and evaluated for the action phase, and accuracy is computed for each action category. The final result is reported as an average over multiple action categories.

5.2.2. Phase Progression

Phase progression represents the extent of an action phase completed at a current frame position. On top the frozen SSL video backbone, a linear layer is trained for regression.

5.2.3. Kendall Tau

Kendall Tau assesses the ordinal knowledge in the learned framewise representation. It is computed across every pair of testing videos by sampling two frames from the first video and retrieving the corresponding nearest frames in the second video to verify if their order is preserved.

5.2.4. Observations

We evaluate both regular video SSL techniques such as Shuffle & Learn (SaL) [13], TCN [164], and TCLR [94] and expert video SSL methods dedicated for the action phases such as CARL [45] and VSP [46] as shown in Table 1. Our first observation is that the video SSL method dedicated to action phases by learning temporal coherent framewise representations, significantly outperform regular videoSSL methods which are proposed mainly for semantic downstream tasks such as action recognition.

One possible drawback of the action phase related evaluation is that the test datasets are small scale (<1k videos) and contains only a few action categories from sports(<14 action classes). This makes the evaluation less reliable and less applicable to the more general action classes such as Kinetics700 [153]. Another potential future direction is to study the shift from ego-centric to third-person viewpoints during action phases, as recently explored in [142].

Table 1. Performance on Phase-related downstream tasks on PennAction and Pouring datasets

Method	Phase Classification		Phase Progression		Kendall Tau	
	PennAction	Pouring	PennAction	Pouring	PennAction	Pouring
SaL [13]	68.2	-	39.0	-	47.4	-
TCN [164]	68.1	89.5	38.3	80.4	54.2	85.2
TCLR [94]	79.9	-	-	-	82.0	-
CARL [45]	93.1	93.7	91.8	93.5	98.5	99.2
VSP [46]	93.1	93.9	92.3	94.2	98.6	99.0

5.3. Frame-to-Frame Temporal Correspondence Based Tasks

5.3.1. Video Object Segmentation (VOS):

VOS addresses the tracking-like correspondence problem where the ground-truth object mask of the first frame is provided, and the model is required to propagate it through subsequent frames. The standard protocol [53] prohibits finetuning any layer of the SSL pretrained model. Evaluation is conducted on the test videos of the DAVIS2017 dataset [165], which features multiple objects simultaneously undergoing deformations, scale changes, and occlusions. Table 2 shows the results of this evaluation.

 doi:10.20944/preprints202408.0133.v1

Table 2. Performance on Temporal Correspondence tasks: video object segmentation on DAVIS and Pose propagation on JHMDB

Method	Venue	DAVIS			JHMDB	
		J&Fm	Jm	Fm	PCK@0.1	PCK@0.2
VFS [53]	ICCV-2021	68.9	66.5	71.3	60.9	80.7
UVC [166]	NeurIPS-2019	56.3	54.5	58.1	56.0	76.6
CRW [51]	NeurIPS-2020	67.6	64.8	70.2	59.3	80.3
SimSiam [79]	CVPR-2021	66.3	64.5	68.2	58.4	77.5
MoCo [87]	MoCo-2020	65.4	63.2	67.6	60.4	79.3
VINCE [167]	arXiv-2020	65.6	63.4	67.8	58.2	76.3
TimeCycle [168]	CVPR-2019	40.7	41.9	39.4	57.7	78.5
MCRW [52]	CVPR-2022	57.9	-	-	62.6	80.9
FGVC [169]	ICCV-2023	77.4	70.5	74.4	66.8	-
NMS [33]	AAAI-2024	62.1	60.5	63.6	43.1	69.7
ST-MAE [66]	Neurips-2022	53.5	52.6	54.4	-	-
VideoMAE [65]	Neurips-2022	53.8	53.2	54.4	36.5	62.1
MotionMAE [68]	arXiv-2022	56.8	55.8	57.8	-	-
SVT [40]	CVPR-2022	48.5	46.8	50.1	35.3	62.66
StT [54]	CVPR-2023	74.1	71.1	77.1	63.1	82.9

5.3.2. Pose Propagation

Pose propagation addresses the temporal correspondence of key points in human poses. The keypoints for 13 joints are provided in the first frame and are tracked using the features of the frozen SSL model. Evaluation is performed on the JHMDB dataset [170], which includes videos from 21 action categories. Table 2 reports the results of this evaluation.

5.3.3. Gait Recognition

Human gait recognition is a biometrics challenge where actors need to be identified based on their gait patterns. The standard gait recognition employs a zero-shot retrieval-based protocol, wherein the model is pretrained on unlabeled video data and then frozen. Subsequently, the nearest neighbor of the query video is identified from the gallery videos (search space). The evaluation is most commonly conducted on the CASIA-B dataset [171], which includes three different splits to indicate varying aspects: normal, viewpoint change, and cloth change. The evaluation of various gait-specific SSL methods and regular video SSL methods is presented in Table 3.

Table 3. Performance on Gait recognition on CASIA-B

Method	Venue	Pretraining Dataset	NM	BG	CL	Mean
GaitSSB [110]	T-PAMI (2023)	GaitLU-1M (1 million videos)	83.30	75.60	**28.70**	62.53
ρBYOL [86]	CVPR (2022)	Kinetics400 (160k videos)	90.65	80.51	28.59	66.58
VideoMAE [65]	Neurips (2022)	Kinetics400 (160k videos)	65.30	57.21	21.40	47.97
NMS [33]	AAAI (2024)	Kinetics400 (160k videos)	**98.60**	**92.57**	28.66	**73.28**

5.3.4. Observations

Our first observation from Table 2 is that the expert videoSSL methods for temporal correspondence significantly outperform the regular action-realted VideoSSL methods. It is also noteworthy that in action-realted videoSSL methods, NMS [33], which relies on the frame-level temporal correspondence-based pretext tasks learning objective, shows commendable performance on VOS and Pose Propagation. Overall, Table 2 indicates that video SSL methods [33,65,86], successful in learning higher-level video semantics such as actions, do not perform well in addressing the low-level temporal correspondence tasks.

Our second observation from Table 3 is that compared to regular videoSSL methods, the expert gait method GaitSSB [110] performs slightly better in the cloth-changing splits, however, it does not perform well on the regular gait splits. The main reason behind this is that [110] solely utilizes precomputed

 doi:10.20944/preprints202408.0133.v1

silhouettes from RGB videos, where its performance heavily depends on the segmentation model used for silhouette computation. Our observations indicate a lack of gait-specific videoSSL methods that outperform regular videoSSL in all cases of appearance and viewpoint. This opens an interesting direction to utilize both silhouette and RGB streams and learn a joint embedding which could aid in learning gait patterns in different scenarios.

5.4. Optical Flow Estimation Related Downstream Task

Estimating optical flow is a fundamental problem in computer vision, crucial for tasks such as visual odometry, depth estimation, and object tracking. For the evaluation, the optical flow estimated by the SSL representation is compared with the ground truth, and results are reported using the standard average endpoint error (EPE). We examine results on two commonly used datasets for this task: Sintel [172] (Clean and Final splits) and KITTI 2015 [173].

5.4.1. Observations

Our first observation from Table 4 is that, although frame-feature correspondence-based methods such as MCRW and multi-task optimization methods like MC-JEPA are capable of performing optical flow estimation, their object-semantic context does not aid in estimating optical flow as effectively as the expert SSL methods for optical flow, such as SMURF and SAFTNet. Furthermore, the joint optimization of both optical-flow and semantic-flow in the form of multi-tasking in MC-JEPA only provides minimal improvement compared to the dedicated semantic-flow-only based method MCRW. This suggests that there is a research opportunity to combine these two objectives more effectively than merely through multi-tasking.

Table 4. Performance on optical flow estimation on Sintel Clean, Sintel Final and KITTI 2015 datasets

Method	Sintel Clean	Sintel Final	KITTI 2015
	AEPE ↓	AEPE ↓	AEPE ↓
MCRW [52]	5.68	6.72	11.67
MC-JEPA [146]	5.01	6.12	11.33
SMURF [71]	3.15	4.18	6.83
SAFTNet [73]	2.44	3.89	-

5.5. Robustness

In order to study the robustness of the VideoSSL model we consider the studies on distribution shifts [93] and input perturbation [33]. In both studies first, videoSSL models pretrained on unlabeled Kinetics dataset are considered and evaluated on the different downstream datasets.

5.5.1. Distribution Shifts

We consider mainly two distribution shifts from [93] here (1) Viewpoint shift and (2) Actor shift. Viewpoint shift refers to the change in the viewpoint or perspective of the camera that captures the scene. Here for the test set we consider only the egocentric view and utilize CharadesEgo [174] dataset. For the actor shift, which occurs when the 'type' of actors changes between training and test sets. These actor-type shifts can include human-to-animal shifts or synthetic-to-real shifts, for which ActorShift [175] and Sims4action [176] datasets are used. Results on such distribution shifts are shown in Figure 9 for both linear and fine-tuned settings. In both cases, we can see that all non-generative methods perform equally better.

 doi:10.20944/preprints202408.0133.v1

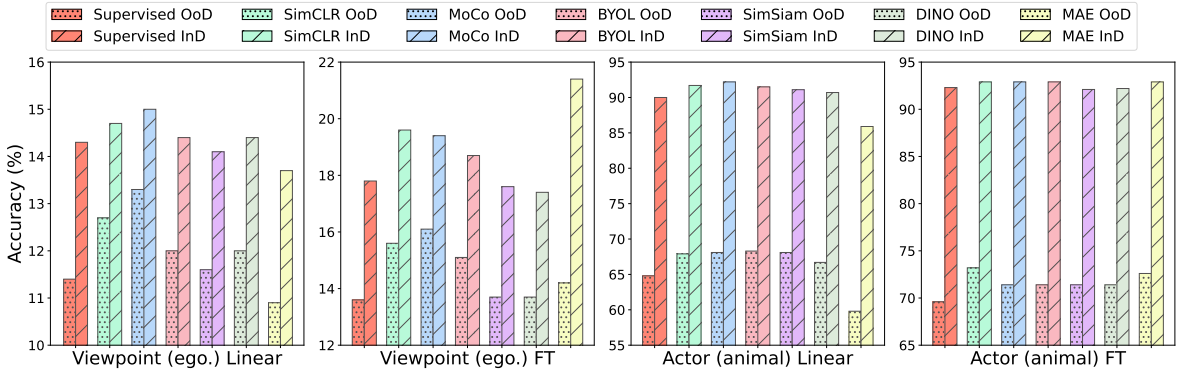


Figure 9. Robustness Analysis with **Distribution shifts**

5.5.2. Input Perturbations

Input perturbation means the input video is corrupted with various frame-independent transformations such as translations, JPEG compression, Gaussian noise, etc. We consider the perturbation study from [33], where performance on zero-shot action retrieval on HMDB51 is considered for various VideoSSL method. The results are shown in Figure 10, where we can see that the non-generative approaches based on the transformer i.e. SVT [40] and NMS [33] provide the most robust representations.

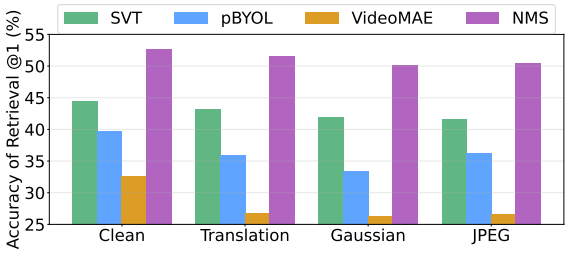


Figure 10. Robust Analysis with **Input Perturbation** on Action retrieval performance on HMDB51 dataset

5.5.3. Observations

Although the self-supervised video representations are expected to provide robustness, there are only a few works have studied them. The current studies only covers the action related downstream tasks and lacks a more in-depth analysis on different downstream tasks. In the future, it is also interesting to study the robustness of the self-supervised video representation in case of adversarial attacks. Currently, to the best of our knowledge, there are no videoSSL works which explicitly focus on increasing the robustness during pretraining so it provides a crucial future direction in videoSSL.

5.6. Overall Experimental Summary and Future Directions

Action Recognition

With varying levels of finetuning, different SSL methods significantly outperform each other. For instance, masking-based methods [65,67] perform the best when fully-finetuned, whereas contrastive learning-based methods [33,40] excel in linear classification or zero-shot video-to-video retrieval. However, current action recognition benchmarks focus on coarse-grained actions and lack a comprehensive benchmark that includes fine-grained action recognition datasets with action categories from the same environment [156,157].

Video Attribute Recognition

Current videoSSL methods initialized with image-based SSL weights achieve better holistic performance for multi-label video classification of attributes such as scene, object, and actions. This

 doi:10.20944/preprints202408.0133.v1

aspect is underexplored in evaluating videoSSL methods. Future robust evaluations should include more diverse multi-class video datasets such as Coin [160] and APPROVE [161].

Phase-Related Tasks

VideoSSL methods [45,46] designed for temporal alignment objectives significantly outperform regular videoSSL methods [40,94] in high-level video understanding. Future evaluations could consider a more practical setup of aligning phases across egocentric to exocentric camera viewpoints [142] with a higher number of action categories instead of just the current 14 categories covered in the existing protocol for more reliable evaluation.

Temporal Correspondence Tasks

Expert video self-supervised learning (videoSSL) methods specializing in temporal correspondence outperform general action-related videoSSL methods in both VOS and Pose Propagation. Although some general videoSSL methods like [33] attempt to learn temporal correspondence through temporal pretext tasks, they do not match the performance of expert methods.

Optical Flow Estimation

In the optical flow estimation task, expert methods like SMURF [71] and SAFTNet [73] outperform regular videoSSL methods like MCRW [52]. While MC-JEPA [146] attempts to show results on both tasks, it significantly lacks performance in optical flow estimation compared to expert flow methods. Future directions could involve combining both optical flow and semantic RGB flow in a self-supervised objective in an effective way with cross interactions rather than simple multi-tasking as in [146].

6. Conclusion

This paper offers the first thorough survey that integrates the full spectrum of video self-supervised learning (VideoSSL), ranging from high-level recognition to low-level temporal correspondence tasks. By conceptually connecting different SSL objectives and providing an in-depth categorization, our work elucidates the varied capabilities and benefits of VideoSSL across multiple task families through detailed evaluations. Our evaluations reveal that although there are some initial attempts to solve more than one family of tasks, they are not good enough and are still mainly inclined towards one family.

Moreover, we outline various challenges associated with videoSSL and acknowledge the strengths of existing research that has established preliminary foundations in this field. We aim for this comprehensive survey to serve as a pivotal resource that advances understanding and inspires further exploration in the rich and diverse domain of VideoSSL.

Acknowledgments: This research was supported by the joint grant P007 from Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI) and the Weizmann Institute of Science (WIS). The authors would like to express their sincere gratitude for this generous support, which made the study possible.

References

1. Hara, K.; Kataoka, H.; Satoh, Y. Learning spatio-temporal features with 3d residual networks for action recognition. Proceedings of the IEEE international conference on computer vision workshops, 2017, pp. 3154–3160.
2. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.
3. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

 doi:10.20944/preprints202408.0133.v1

4. Xie, S.; Sun, C.; Huang, J.; Tu, Z.; Murphy, K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. Proceedings of the European conference on computer vision (ECCV), 2018, pp. 305–321.
5. Neimark, D.; Bar, O.; Zohar, M.; Asselmann, D. Video transformer network. Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 3163–3172.
6. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. Vivit: A video vision transformer. Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 6836–6846.
7. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video swin transformer. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 3202–3211.
8. Girdhar, R.; Carreira, J.; Doersch, C.; Zisserman, A. Video action transformer network. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 244–253.
9. Zhao, H.; Torralba, A.; Torresani, L.; Yan, Z. Hacs: Human action clips and segments dataset for recognition and temporal localization. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8668–8678.
10. Diba, A.; Fayyaz, M.; Sharma, V.; Paluri, M.; Gall, J.; Stiefelhagen, R.; Van Gool, L. Large scale holistic video understanding. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer, 2020, pp. 593–610.
11. Schiappa, M.C.; Rawat, Y.S.; Shah, M. Self-supervised learning for videos: A survey. *ACM Computing Surveys* **2022**.
12. Bardes, A.; Garrido, Q.; Ponce, J.; Chen, X.; Rabbat, M.; LeCun, Y.; Assran, M.; Ballas, N. V-JEPA: Latent Video Prediction for Visual Representation Learning **2023**.
13. Misra, I.; Zitnick, C.L.; Hebert, M. Shuffle and learn: unsupervised learning using temporal order verification. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, 2016, pp. 527–544.
14. Fernando, B.; Bilen, H.; Gavves, E.; Gould, S. Self-supervised video representation learning with odd-one-out networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3636–3645.
15. Lee, H.Y.; Huang, J.B.; Singh, M.; Yang, M.H. Unsupervised representation learning by sorting sequences. Proceedings of the IEEE international conference on computer vision, 2017, pp. 667–676.
16. Wei, D.; Lim, J.J.; Zisserman, A.; Freeman, W.T. Learning and using the arrow of time. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8052–8060.
17. Xu, D.; Xiao, J.; Zhao, Z.; Shao, J.; Xie, D.; Zhuang, Y. Self-supervised spatiotemporal learning via video clip order prediction. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10334–10343.
18. Xue, F.; Ji, H.; Zhang, W.; Cao, Y. Self-supervised video representation learning by maximizing mutual information. *Signal processing: Image communication* **2020**, *88*, 115967.
19. Luo, D.; Liu, C.; Zhou, Y.; Yang, D.; Ma, C.; Ye, Q.; Wang, W. Video cloze procedure for self-supervised spatio-temporal learning. Proceedings of the AAAI conference on artificial intelligence, 2020, Vol. 34, pp. 11701–11708.
20. Hu, K.; Shao, J.; Liu, Y.; Raj, B.; Savvides, M.; Shen, Z. Contrast and order representations for video self-supervised learning. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7939–7949.
21. Yao, T.; Zhang, Y.; Qiu, Z.; Pan, Y.; Mei, T. Seco: Exploring sequence supervision for unsupervised representation learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, Vol. 35, pp. 10656–10664.
22. Guo, S.; Xiong, Z.; Zhong, Y.; Wang, L.; Guo, X.; Han, B.; Huang, W. Cross-architecture self-supervised video representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19270–19279.
23. Luo, D.; Zhou, Y.; Fang, B.; Zhou, Y.; Wu, D.; Wang, W. Exploring relations in untrimmed videos for self-supervised learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **2022**, *18*, 1–21.
24. Kumar, V. Unsupervised Learning of Spatio-Temporal Representation with Multi-Task Learning for Video Retrieval. 2022 National Conference on Communications (NCC). IEEE, 2022, pp. 118–123.

 doi:10.20944/preprints202408.0133.v1

25. Dorkenwald, M.; Xiao, F.; Brattoli, B.; Tighe, J.; Modolo, D. Scvrl: Shuffled contrastive video representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4132–4141.
26. Bai, Y.; Fan, H.; Misra, I.; Venkatesh, G.; Lu, Y.; Zhou, Y.; Yu, Q.; Chandra, V.; Yuille, A. Can temporal information help with contrastive self-supervised learning? *arXiv preprint arXiv:2011.13046* **2020**.
27. Liu, Y.; Wang, K.; Liu, L.; Lan, H.; Lin, L. Tcgl: Temporal contrastive graph for self-supervised video representation learning. *IEEE Transactions on Image Processing* **2022**, *31*, 1978–1993.
28. Jenni, S.; Jin, H. Time-equivariant contrastive video representation learning. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9970–9980.
29. Cho, H.; Kim, T.; Chang, H.J.; Hwang, W. Self-supervised visual learning by variable playback speeds prediction of a video. *IEEE Access* **2021**, *9*, 79562–79571.
30. Wang, J.; Jiao, J.; Liu, Y.H. Self-supervised video representation learning by pace prediction. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. Springer, 2020, pp. 504–521.
31. Benaim, S.; Ephrat, A.; Lang, O.; Mosseri, I.; Freeman, W.T.; Rubinstein, M.; Irani, M.; Dekel, T. Speednet: Learning the speediness in videos. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9922–9931.
32. Chen, P.; Huang, D.; He, D.; Long, X.; Zeng, R.; Wen, S.; Tan, M.; Gan, C. Rspnet: Relative speed perception for unsupervised video representation learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, Vol. 35, pp. 1045–1053.
33. Dave, I.R.; Jenni, S.; Shah, M. No More Shortcuts: Realizing the Potential of Temporal Self-Supervision. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 1481–1491.
34. Jenni, S.; Meishvili, G.; Favaro, P. Video representation learning by recognizing temporal transformations. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16. Springer, 2020, pp. 425–442.
35. Yang, C.; Xu, Y.; Dai, B.; Zhou, B. Video representation learning with visual tempo consistency. *arXiv preprint arXiv:2006.15489* **2020**.
36. Behrmann, N.; Fayyaz, M.; Gall, J.; Noroozi, M. Long short view feature decomposition via contrastive video representation learning. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9244–9253.
37. Recasens, A.; Luc, P.; Alayrac, J.B.; Wang, L.; Strub, F.; Tallec, C.; Malinowski, M.; Pătrăucean, V.; Altché, F.; Valko, M.; others. Broaden your views for self-supervised video learning. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1255–1265.
38. Fang, B.; Wu, W.; Liu, C.; Zhou, Y.; He, D.; Wang, W. Mamico: Macro-to-micro semantic correspondence for self-supervised video representation learning. Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1348–1357.
39. Liu, C.; Yao, Y.; Luo, D.; Zhou, Y.; Ye, Q. Self-supervised motion perception for spatiotemporal representation learning. *IEEE Transactions on Neural Networks and Learning Systems* **2022**.
40. Ranasinghe, K.; Naseer, M.; Khan, S.; Khan, F.S.; Ryoo, M.S. Self-supervised video transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2874–2884.
41. Qian, R.; Li, Y.; Yuan, L.; Gong, B.; Liu, T.; Brown, M.; Belongie, S.J.; Yang, M.H.; Adam, H.; Cui, Y. On Temporal Granularity in Self-Supervised Video Representation Learning. BMVC, 2022, p. 541.
42. Jeong, S.Y.; Kim, H.J.; Oh, M.S.; Lee, G.H.; Lee, S.W. Temporal-Invariant Video Representation Learning with Dynamic Temporal Resolutions. 2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2022, pp. 1–8.
43. Yao, Y.; Liu, C.; Luo, D.; Zhou, Y.; Ye, Q. Video playback rate perception for self-supervised spatio-temporal representation learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 6548–6557.
44. Knights, J.; Harwood, B.; Ward, D.; Vanderkop, A.; Mackenzie-Ross, O.; Moghadam, P. Temporally coherent embeddings for self-supervised video representation learning. 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 8914–8921.

45. Chen, M.; Wei, F.; Li, C.; Cai, D. Frame-wise Action Representations for Long Videos via Sequence Contrastive Learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13801–13810.
46. Zhang, H.; Liu, D.; Zheng, Q.; Su, B. Modeling video as stochastic processes for fine-grained video representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2225–2234.
47. Dwibedi, D.; Aytar, Y.; Tompson, J.; Sermanet, P.; Zisserman, A. Temporal cycle-consistency learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1801–1810.
48. Hadji, I.; Derpanis, K.G.; Jepson, A.D. Representation learning via global temporal alignment and cycle-consistency. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11068–11077.
49. Haresh, S.; Kumar, S.; Coskun, H.; Syed, S.N.; Konin, A.; Zia, Z.; Tran, Q.H. Learning by aligning videos in time. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5548–5558.
50. Wills, J.; Agarwal, S.; Belongie, S. What went where. Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; IEEE Computer Society: USA, 2003; CVPR'03, p. 37–44.
51. Jabri, A.; Owens, A.; Efros, A. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems* **2020**, *33*, 19545–19560.
52. Bian, Z.; Jabri, A.; Efros, A.A.; Owens, A. Learning pixel trajectories with multiscale contrastive random walks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 6508–6519.
53. Xu, J.; Wang, X. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10075–10085.
54. Li, R.; Liu, D. Spatial-then-temporal self-supervised learning for video correspondence. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2279–2288.
55. Noroozi, M.; Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. European conference on computer vision. Springer, 2016, pp. 69–84.
56. Ahsan, U.; Madhok, R.; Essa, I. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019, pp. 179–189.
57. Gidaris, S.; Singh, P.; Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728* **2018**.
58. Jing, L.; Yang, X.; Liu, J.; Tian, Y. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387* **2018**.
59. Geng, S.; Zhao, S.; Liu, H. Video representation learning by identifying spatio-temporal transformations. *Applied Intelligence* **2022**, pp. 1–10.
60. Zhang, Y.; Po, L.M.; Xu, X.; Liu, M.; Wang, Y.; Ou, W.; Zhao, Y.; Yu, W.Y. Contrastive spatio-temporal pretext learning for self-supervised video representation. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 3380–3389.
61. Wang, R.; Chen, D.; Wu, Z.; Chen, Y.; Dai, X.; Liu, M.; Yuan, L.; Jiang, Y.G. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6312–6322.
62. Buchler, U.; Brattoli, B.; Ommer, B. Improving spatiotemporal self-supervision by deep reinforcement learning. Proceedings of the European conference on computer vision (ECCV), 2018, pp. 770–786.
63. Kim, D.; Cho, D.; Kweon, I.S. Self-supervised video representation learning with space-time cubic puzzles. Proceedings of the AAAI conference on artificial intelligence, 2019, Vol. 33, pp. 8545–8552.
64. Zhang, Y.; Zhang, H.; Wu, G.; Li, J. Spatio-temporal self-supervision enhanced transformer networks for action recognition. 2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022, pp. 1–6.
65. Tong, Z.; Song, Y.; Wang, J.; Wang, L. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. Advances in Neural Information Processing Systems, 2022.

 doi:10.20944/preprints202408.0133.v1

66. Feichtenhofer, C.; Fan, H.; Li, Y.; He, K. Masked Autoencoders As Spatiotemporal Learners. Advances in Neural Information Processing Systems, 2022.
67. Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Wang, Y.; Qiao, Y. Videomae v2: Scaling video masked autoencoders with dual masking. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14549–14560.
68. Yang, H.; Huang, D.; Wen, B.; Wu, J.; Yao, H.; Jiang, Y.; Zhu, X.; Yuan, Z. Self-supervised Video Representation Learning with Motion-Aware Masked Autoencoders. *arXiv preprint arXiv:2210.04154* **2022**.
69. Song, Y.; Yang, M.; Wu, W.; He, D.; Li, F.; Wang, J. It Takes Two: Masked Appearance-Motion Modeling for Self-supervised Video Transformer Pre-training. *arXiv preprint arXiv:2210.05234* **2022**.
70. Sun, X.; Chen, P.; Chen, L.; Li, C.; Li, T.H.; Tan, M.; Gan, C. Masked Motion Encoding for Self-Supervised Video Representation Learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2235–2245.
71. Stone, A.; Maurer, D.; Ayvaci, A.; Angelova, A.; Jonschkowski, R. Smurf: Self-teaching multi-frame unsupervised raft with full-image warping. Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2021, pp. 3887–3896.
72. Teed, Z.; Deng, J. Raft: Recurrent all-pairs field transforms for optical flow. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer, 2020, pp. 402–419.
73. Sun, Z.; Luo, Z.; Nishida, S. Decoupled spatiotemporal adaptive fusion network for self-supervised motion estimation. *Neurocomputing* **2023**, *534*, 133–146.
74. Gan, C.; Gong, B.; Liu, K.; Su, H.; Guibas, L.J. Geometry guided convolutional neural networks for self-supervised video representation learning. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5589–5597.
75. Sriram, A.; Gaidon, A.; Wu, J.; Niebles, J.C.; Fei-Fei, L.; Adeli, E. HomE: Homography-Equivariant Video Representation Learning. *arXiv preprint arXiv:2306.01623* **2023**.
76. Das, S.; Ryoo, M.S. ViewCLR: Learning Self-supervised Video Representation for Unseen Viewpoints. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 5573–5583.
77. Gutmann, M.; Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.
78. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; others. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* **2020**, *33*, 21271–21284.
79. Chen, X.; He, K. Exploring simple siamese representation learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 15750–15758.
80. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* **2018**.
81. Han, T.; Xie, W.; Zisserman, A. Video representation learning by dense predictive coding. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.
82. Han, T.; Xie, W.; Zisserman, A. Memory-augmented dense predictive coding for video representation learning. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. Springer, 2020, pp. 312–329.
83. Lorre, G.; Rabarisoa, J.; Orcesi, A.; Ainouz, S.; Canu, S. Temporal Contrastive Pretraining for Video Action Recognition. The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 662–670.
84. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. International conference on machine learning. PMLR, 2020, pp. 1597–1607.
85. Qian, R.; Meng, T.; Gong, B.; Yang, M.H.; Wang, H.; Belongie, S.; Cui, Y. Spatiotemporal contrastive video representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6964–6974.
86. Feichtenhofer, C.; Fan, H.; Xiong, B.; Girshick, R.; He, K. A large-scale study on unsupervised spatiotemporal representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3299–3309.

 doi:10.20944/preprints202408.0133.v1

87. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9729–9738.
88. Pan, T.; Song, Y.; Yang, T.; Jiang, W.; Liu, W. Videomoco: Contrastive video representation learning with temporally adversarial examples. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11205–11214.
89. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems* **2020**, *33*, 9912–9924.
90. Hjelm, R.D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670* **2018**.
91. Bachman, P.; Hjelm, R.D.; Buchwalter, W. Learning representations by maximizing mutual information across views. Advances in Neural Information Processing Systems, 2019, pp. 15535–15545.
92. Devon Hjelm, R.; Bachman, P. Representation Learning with Video Deep InfoMax. *arXiv preprint arXiv:2007.13278* **2020**.
93. Sarkar, P.; Beirami, A.; Etemad, A. Uncovering the Hidden Dynamics of Video Self-supervised Learning under Distribution Shifts. *arXiv preprint arXiv:2306.02014* **2023**.
94. Dave, I.; Gupta, R.; Rizve, M.N.; Shah, M. TCLR: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding* **2022**, *219*, 103406.
95. Wang, J.; Lin, Y.; Ma, A.J.; Yuen, P.C. Self-supervised temporal discriminative learning for video representation learning. *arXiv preprint arXiv:2008.02129* **2020**.
96. Chen, Z.; Lin, K.Y.; Zheng, W.S. Consistent Intra-video Contrastive Learning with Asynchronous Long-term Memory Bank. *IEEE Transactions on Circuits and Systems for Video Technology* **2022**.
97. Tao, L.; Wang, X.; Yamasaki, T. Self-supervised video representation learning using inter-intra contrastive framework. Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2193–2201.
98. Tao, L.; Wang, X.; Yamasaki, T. An improved inter-intra contrastive learning framework on self-supervised video representation. *IEEE Transactions on Circuits and Systems for Video Technology* **2022**, *32*, 5266–5280.
99. Zhu, Y.; Shuai, H.; Liu, G.; Liu, Q. Self-supervised video representation learning using improved instance-wise contrastive learning and deep clustering. *IEEE Transactions on Circuits and Systems for Video Technology* **2022**, *32*, 6741–6752.
100. Khorasgani, S.H.; Chen, Y.; Shkurti, F. Slic: Self-supervised learning with iterative clustering for human action videos. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16091–16101.
101. Miech, A.; Alayrac, J.B.; Smaira, L.; Laptev, I.; Sivic, J.; Zisserman, A. End-to-end learning of visual representations from uncurated instructional videos. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9879–9889.
102. Tokmakov, P.; Hebert, M.; Schmid, C. Unsupervised learning of video representations via dense trajectory clustering. Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer, 2020, pp. 404–421.
103. Zach, C.; Pock, T.; Bischof, H. A duality based approach for realtime tv-l 1 optical flow. Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29. Springer, 2007, pp. 214–223.
104. Wang, J.; Jiao, J.; Bao, L.; He, S.; Liu, Y.; Liu, W. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4006–4015.
105. Wang, J.; Jiao, J.; Bao, L.; He, S.; Liu, W.; Liu, Y.h. Self-supervised video representation learning by uncovering spatio-temporal statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2021**, *44*, 3791–3806.
106. Xiao, F.; Tighe, J.; Modolo, D. Maclr: Motion-aware contrastive learning of representations for videos. European Conference on Computer Vision. Springer, 2022, pp. 353–370.
107. Ni, J.; Zhou, N.; Qin, J.; Wu, Q.; Liu, J.; Li, B.; Huang, D. Motion Sensitive Contrastive Learning for Self-supervised Video Representation. Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV. Springer, 2022, pp. 457–474.

 doi:10.20944/preprints202408.0133.v1

108. Coskun, H.; Zareian, A.; Moore, J.L.; Tombari, F.; Wang, C. GOCA: guided online cluster assignment for self-supervised video representation Learning. Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI. Springer, 2022, pp. 1–22.
109. Han, T.; Xie, W.; Zisserman, A. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems* **2020**, *33*, 5679–5690.
110. Fan, C.; Hou, S.; Wang, J.; Huang, Y.; Yu, S. Learning gait representation from massive unlabelled walking videos: A benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**.
111. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. Springer, 2020, pp. 173–190.
112. Alwassel, H.; Mahajan, D.; Korbar, B.; Torresani, L.; Ghanem, B.; Tran, D. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems* **2020**, *33*, 9758–9770.
113. Kalayeh, M.M.; Kamath, N.; Liu, L.; Chandrashekar, A. Watching too much television is good: Self-supervised audio-visual representation learning from movies and tv shows. *arXiv preprint arXiv:2106.08513* **2021**.
114. Kalayeh, M.M.; Ardeshir, S.; Liu, L.; Kamath, N.; Chandrashekar, A. On Negative Sampling for Audio-Visual Contrastive Learning from Movies. *arXiv preprint arXiv:2205.00073* **2022**.
115. Qing, Z.; Zhang, S.; Huang, Z.; Xu, Y.; Wang, X.; Tang, M.; Gao, C.; Jin, R.; Sang, N. Learning from untrimmed videos: Self-supervised video representation learning with hierarchical consistency. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13821–13831.
116. Qing, Z.; Zhang, S.; Huang, Z.; Xu, Y.; Wang, X.; Gao, C.; Jin, R.; Sang, N. Self-Supervised Learning from Untrimmed Videos via Hierarchical Consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**.
117. Li, R.; Zhou, S.; Liu, D. Learning Fine-Grained Features for Pixel-wise Video Correspondences. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 9632–9641.
118. Thoker, F.M.; Doughty, H.; Snoek, C. Tubelet-Contrastive Self-Supervision for Video-Efficient Generalization. *arXiv preprint arXiv:2303.11003* **2023**.
119. Kong, Q.; Wei, W.; Deng, Z.; Yoshinaga, T.; Murakami, T. Cycle-contrast for self-supervised video representation learning. *Advances in Neural Information Processing Systems* **2020**, *33*, 8089–8100.
120. Diba, A.; Sharma, V.; Safdari, R.; Lotfi, D.; Sarfraz, S.; Stiefelhagen, R.; Van Gool, L. Vi2clr: Video and image for visual contrastive learning of representation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1502–1512.
121. Lin, W.; Mirza, M.J.; Kozinski, M.; Possegger, H.; Kuehne, H.; Bischof, H. Video Test-Time Adaptation for Action Recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22952–22961.
122. Lin, J.; Zhang, R.; Ganz, F.; Han, S.; Zhu, J.Y. Enhancing Unsupervised Video Representation Learning by Decoupling the Scene and the Motion. AAAI, 2021.
123. Wang, J.; Gao, Y.; Li, K.; Lin, Y.; Ma, A.J.; Cheng, H.; Peng, P.; Huang, F.; Ji, R.; Sun, X. Removing the background by adding the background: Towards background robust self-supervised video representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11804–11813.
124. Zhang, M.; Wang, J.; Ma, A.J. Suppressing Static Visual Cues via Normalizing Flows for Self-Supervised Video Representation Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 3300–3308.
125. Tian, F.; Fan, J.; Yu, X.; Du, S.; Song, M.; Zhao, Y. TCVM: Temporal Contrasting Video Montage Framework for Self-supervised Video Representation Learning. Proceedings of the Asian Conference on Computer Vision, 2022, pp. 1539–1555.
126. Ding, S.; Qian, R.; Xiong, H. Dual contrastive learning for spatio-temporal representation. Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5649–5658.
127. Akar, A.; Senturk, U.U.; Ikizler-Cinbis, N. MAC: Mask-Augmentation for Motion-Aware Video Representation Learning. 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022. BMVA Press, 2022.

128. Assefa, M.; Jiang, W.; Gedamu, K.; Yilma, G.; Kumeda, B.; Ayalew, M. Self-Supervised Scene-Debiasing for Video Representation Learning via Background Patching. *IEEE Transactions on Multimedia* **2022**.
129. Kim, J.; Kim, T.; Shim, M.; Han, D.; Wee, D.; Kim, J. Spatiotemporal Augmentation on Selective Frequencies for Video Representation Learning. *arXiv preprint arXiv:2204.03865* **2022**.
130. Chen, B.; Selvaraju, R.R.; Chang, S.F.; Niebles, J.C.; Naik, N. Previts: contrastive pretraining with video tracking supervision. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 1560–1570.
131. Huang, L.; Liu, Y.; Wang, B.; Pan, P.; Xu, Y.; Jin, R. Self-supervised video representation learning by context and motion decoupling. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13886–13895.
132. Ding, S.; Li, M.; Yang, T.; Qian, R.; Xu, H.; Chen, Q.; Wang, J.; Xiong, H. Motion-aware contrastive video representation learning via foreground-background merging. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9716–9726.
133. Liu, J.; Cheng, Y.; Zhang, Y.; Zhao, R.W.; Feng, R. Self-Supervised Video Representation Learning with Motion-Contrastive Perception. 2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022, pp. 1–6.
134. Nie, M.; Quan, Z.; Ding, W.; Yang, W. Enhancing motion visual cues for self-supervised video representation learning. *Engineering Applications of Artificial Intelligence* **2023**, *123*, 106203.
135. Gupta, R.; Akhtar, N.; Mian, A.; Shah, M. Contrastive self-supervised learning leads to higher adversarial susceptibility. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, Vol. 37, pp. 14838–14846.
136. Yu, Y.; Lee, S.; Kim, G.; Song, Y. Self-supervised learning of compressed video representations. International Conference on Learning Representations, 2021.
137. Hwang, S.; Yoon, J.; Lee, Y.; Hwang, S.J. Efficient Video Representation Learning via Masked Video Modeling with Motion-centric Token Selection. *arXiv preprint arXiv:2211.10636* **2022**.
138. Li, Q.; Huang, X.; Wan, Z.; Hu, L.; Wu, S.; Zhang, J.; Shan, S.; Wang, Z. Data-Efficient Masked Video Modeling for Self-supervised Action Recognition. Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 2723–2733.
139. Qing, Z.; Zhang, S.; Huang, Z.; Wang, X.; Wang, Y.; Lv, Y.; Gao, C.; Sang, N. Mar: Masked autoencoders for efficient action recognition. *IEEE Transactions on Multimedia* **2023**.
140. Dave, I.R.; Rizve, M.N.; Chen, C.; Shah, M. Timebalance: Temporally-invariant and temporally-distinctive video representations for semi-supervised action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2341–2352.
141. Escorcia, V.; Guerrero, R.; Zhu, X.; Martinez, B. SOS! Self-supervised Learning over Sets of Handled Objects in Egocentric Action Recognition. Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII. Springer, 2022, pp. 604–620.
142. Xue, Z.S.; Grauman, K. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems* **2024**, *36*.
143. Dave, I.R.; Chen, C.; Shah, M. Spact: Self-supervised privacy preservation for action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20164–20173.
144. Fioresi, J.; Dave, I.R.; Shah, M. Ted-spad: Temporal distinctiveness for self-supervised privacy-preservation for video anomaly detection. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 13598–13609.
145. Rehman, Y.A.U.; Gao, Y.; Shen, J.; de Gusmao, P.P.B.; Lane, N. Federated self-supervised learning for video understanding. European Conference on Computer Vision. Springer, 2022, pp. 506–522.
146. Bardes, A.; Ponce, J.; LeCun, Y. Mc-jepa: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv preprint arXiv:2307.12698* **2023**.
147. Lai, Q.; Zeng, A.; Wang, Y.; Cao, L.; Li, Y.; Xu, Q. Self-supervised Video Representation Learning via Capturing Semantic Changes Indicated by Saccades. *IEEE Transactions on Circuits and Systems for Video Technology* **2023**.
148. Lin, W.; Ding, X.; Huang, Y.; Zeng, H. Self-Supervised Video-Based Action Recognition With Disturbances. *IEEE Transactions on Image Processing* **2023**.

 doi:10.20944/preprints202408.0133.v1

149. Qian, R.; Ding, S.; Liu, X.; Lin, D. Static and Dynamic Concepts for Self-supervised Video Representation Learning. Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI. Springer, 2022, pp. 145–164.
150. Lin, W.; Liu, X.; Zhuang, Y.; Ding, X.; Tu, X.; Huang, Y.; Zeng, H. Unsupervised Video-based Action Recognition With Imagining Motion And Perceiving Appearance. *IEEE Transactions on Circuits and Systems for Video Technology* **2022**.
151. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* **2012**.
152. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: a large video database for human motion recognition. 2011 International conference on computer vision. IEEE, 2011, pp. 2556–2563.
153. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; others. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* **2017**.
154. Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; others. The" something something" video database for learning and evaluating visual common sense. Proceedings of the IEEE international conference on computer vision, 2017, pp. 5842–5850.
155. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1010–1019.
156. Shao, D.; Zhao, Y.; Dai, B.; Lin, D. Finegym: A hierarchical video dataset for fine-grained action understanding. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2616–2625.
157. Li, Y.; Li, Y.; Vasconcelos, N. Resound: Towards action recognition without representation bias. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 513–528.
158. Thoker, F.M.; Doughty, H.; Bagad, P.; Snoek, C.G. How Severe Is Benchmark-Sensitivity in Video Self-supervised Learning? Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV. Springer, 2022, pp. 632–652.
159. Diba, A.; Fayyaz, M.; Sharma, V.; Paluri, M.; Gall, J.; Stiefelhagen, R.; Van Gool, L. Large Scale Holistic Video Understanding. European Conference on Computer Vision. Springer, 2020, pp. 593–610.
160. Tang, Y.; Ding, D.; Rao, Y.; Zheng, Y.; Zhang, D.; Zhao, L.; Lu, J.; Zhou, J. Coin: A large-scale dataset for comprehensive instructional video analysis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1207–1216.
161. Gupta, R.; Roy, A.; Christensen, C.; Kim, S.; Gerard, S.; Cincebeaux, M.; Divakaran, A.; Grindal, T.; Shah, M. Class prototypes based contrastive learning for classifying multi-label and fine-grained educational videos. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 19923–19933.
162. Zhang, W.; Zhu, M.; Derpanis, K.G. From actemes to action: A strongly-supervised representation for detailed action understanding. Proceedings of the IEEE international conference on computer vision, 2013, pp. 2248–2255.
163. Sermanet, P.; Xu, K.; Levine, S. Unsupervised Perceptual Rewards for Imitation Learning. Proceedings of Robotics: Science and Systems, 2017.
164. Sermanet, P.; Lynch, C.; Hsu, J.; Levine, S. Time-contrastive networks: Self-supervised learning from multi-view observation. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2017, pp. 486–487.
165. Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; Van Gool, L. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* **2017**.
166. Li, X.; Liu, S.; De Mello, S.; Wang, X.; Kautz, J.; Yang, M.H. Joint-task self-supervised learning for temporal correspondence. *Advances in Neural Information Processing Systems* **2019**, *32*.
167. Gordon, D.; Ehsani, K.; Fox, D.; Farhadi, A. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990* **2020**.
168. Wang, X.; Jabri, A.; Efros, A.A. Learning correspondence from the cycle-consistency of time. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2566–2576.
169. Li, R.; Zhou, S.; Liu, D. Learning Fine-Grained Features for Pixel-wise Video Correspondences. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 9632–9641.

 doi:10.20944/preprints202408.0133.v1

170. Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; Black, M.J. Towards understanding action recognition. International Conf. on Computer Vision (ICCV), 2013, pp. 3192–3199.
171. Yu, S.; Tan, D.; Tan, T. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. 18th International Conference on Pattern Recognition (ICPR'06). IEEE, 2006, Vol. 4, pp. 441–444.
172. Butler, D.J.; Wulff, J.; Stanley, G.B.; Black, M.J. A naturalistic open source movie for optical flow evaluation. Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12. Springer, 2012, pp. 611–625.
173. Menze, M.; Heipke, C.; Geiger, A. Joint 3d estimation of vehicles and scene flow. *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences* **2015**, *2*, 427–434.
174. Sigurdsson, G.A.; Gupta, A.; Schmid, C.; Farhadi, A.; Alahari, K. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626* **2018**.
175. Zhang, Y.; Doughty, H.; Shao, L.; Snoek, C.G. Audio-adaptive activity recognition across video domains. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13791–13800.
176. Roitberg, A.; Schneider, D.; Djamal, A.; Seibold, C.; Reiß, S.; Stiefelhagen, R. Let's play for action: Recognizing activities of daily living by learning from life simulation video games. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2021, pp. 8563–8569.