Scene Consistency Representation Learning for Video Scene Segmentation

Haoqian Wu[1,2,3,4*], Keyu Chen[2*], Yanan Luo[2], Ruizhi Qiao[2], Bo Ren[2],
Haozhe Liu[1,3,4,5], Weicheng Xie[1,3,4†], Linlin Shen[1,3,4]
[1] Computer Vision Institute, Shenzhen University [2] Tencent YouTu Lab
[3] Shenzhen Institute of Artificial Intelligence and Robotics for Society
[4] Guangdong Key Laboratory of Intelligent Information Processing [5] KAUST
wuhaoqian2019@email.szu.edu.cn {yolochen, ruizhiqiao, timren}@tencent.com
luoyanan93@gmail.com haozhe.liu@kaust.edu.sa {wcxie, llshen}@szu.edu.cn

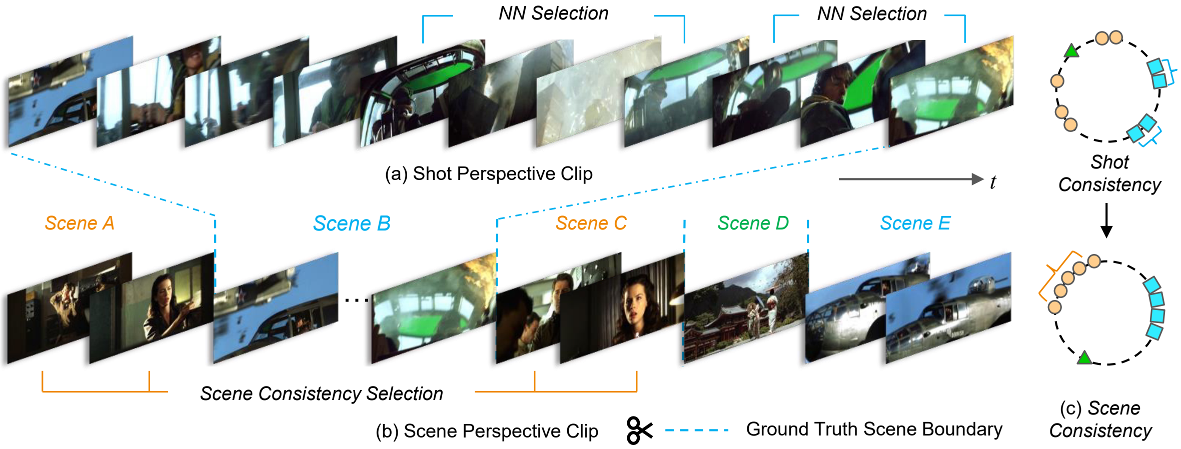
NN Selection
NN Selection
(a) Shot Perspective Clip
t
Shot Consistency
Scene A
Scene B
Scene C
Scene D
Scene E
Scene Consistency Selection
(b) Scene Perspective Clip
Ground Truth Scene Boundary
(c) Scene Consistency

Figure 1. **An illustration of representation learning methods from the shot-to-scene perspective**. Several continuous shots are shown in Fig. (a), where existing SSL approaches obtain positive pairs from the adjacent shots (*e.g.*, by performing *Nearest Neighbor (NN) Selection* [1]). While we propose to look further for scenes that are often crossed over, as *Scene A/C* and *Scene B/E* shown in Fig. (b), where positive samples are explored in a broader region and the shots are clustered to the same scene in the feature representation space, *i.e.*, Fig. (c). Best viewed in color.

Abstract

A long-term video, such as a movie or TV show, is composed of various scenes, each of which represents a series of shots sharing the same semantic story. Spotting the correct scene boundary from the long-term video is a challenging task, since a model must understand the storyline of the video to figure out where a scene starts and ends. To this end, we propose an effective Self-Supervised Learning (SSL) framework to learn better shot representations from unlabeled long-term videos. More specifically, we present an SSL scheme to achieve scene consistency, while exploring considerable data augmentation and shuffling methods to boost the model generalizability. Instead of explicitly learning the scene boundary features as in the previous methods, we introduce a vanilla temporal model with less inductive bias to verify the quality of the shot features. Our method achieves the state-of-the-art performance on the task of Video Scene Segmentation. Additionally, we suggest a more fair and reasonable benchmark to evaluate the performance of Video Scene Segmentation methods. The code is made available.[1]

1. Introduction

In the process of video creation, to make the story more compelling, the editor will use various editing techniques, such as montage, one shot to the end, *etc.* Quickly switching between stories and scenes makes the movie plot tighter, *e.g.* inserting outdoor battle scenes into indoor dialogue scenes, as shown in Fig. 1 (b), making the scene tran-

*Equal Contribution †Corresponding Author

[1] https://github.com/TencentYoutuResearch/SceneSegmentation-SCRL

sition more intriguing and unpredictable, thus the task of *Video Scene Segmentation* turns out to be rather challenging. Hence, it is essential to understand the high-level semantic information of each scene in the long-term video.

There has been extensive studies dealing with video understanding tasks on datasets where the individual video clip is typically short, while requiring a lot of labor to segment uncurated videos into short videos by category. Although some studies focus on splitting the long video into smaller segments, *e.g.*, the methods of *Action Spotting* [2–5] aim to locate the positions of the beginning and ending of the action, however, they are the category-aware approaches. By contrast, *Video Scene Segmentation* is a category-agnostic task that only the scene boundary label is available, and it's very confusing to classify a scene fragment taxonomically.

Since a long-term video is inherently structured in a specific way, a sequence of frames can be divided into **shots** or **scenes** in terms of the granularity of semantics [6] [7]. More specifically, a **shot** contains only continuous frames taken by the camera without interruption, and a **scene** is composed of successive shots and describes the same short story. For detecting shot boundaries, [8] [7] split a video into many separate shots using lower-level visual context. Based on this, many mainstream approaches of *Video Scene Segmentation* [9] [10] [6] [1] determine scene boundaries by exploring semantic correlations among the adjacent shots.

While computer vision tasks suffer from the high cost of manual annotation, Self-Supervised Learning (SSL) based methods [11–18] are proposed to train a general feature extractor using unlabeled data. By leveraging a small amount of annotated data for training, these SSL methods can achieve appealing feature representation to even rival some supervised learning methods. For *Video Scene Segmentation*, [1] proposes to narrow the feature representation distance of the most similar shot pair in a local region, it significantly surpasses the supervised learning method [6] by employing a mere MLP classifier. However, in current SSL methods on the task of *Video Scene Segmentation*, the strategy of positive sample selection, pretraining protocol, evaluation metric and downstream model are not well discussed or addressed.

To achieve this goal, we propose a self-supervised learning scheme to learn better representations, as well as the evaluation metric for the task of *Video Scene Segmentation*. The contributions of this paper are summarized as follows:

- A representation learning scheme based on Scene Consistency is proposed to obtain better shot representations on the unlabeled long-term video.
- A simple yet effective temporal model with less inductive bias is proposed to assess the quality of the shot representation for the downstream *Video Scene Segmentation* task.
- A benchmark that is more fair and reasonable is introduced for both pretraining and evaluation. More importantly, the proposed method outperforms the state-of-the-art methods under all the protocols, and can significantly improve the performance of existing supervised methods without bells and whistles.

2. Related Work

Self-Supervised Learning in Images and Videos. To address the problems of the insufficient and expensive manual annotation, many approaches explore the inherent knowledge in unlabeled data by designing a lot of pretext tasks, including predicting the transformations of images, *e.g.*, image rotation [20], inpainting [21], colorizing [22], jigsaw [23], etc. In short, these Self-Supervised Learning (SSL) methods use the information explored from the data themselves for the supervision. Recently, [11–18] introduce the contrastive similarity metrics to learn invariant feature representation of various views augmented from the original image, where strong data augmentations [13] are frequently used in image-level SSL methods to improve the robustness of the learned representations. From another aspect, by finetuning the model with a small amount of labeled data, SSL methods can achieve competitive performance compared with supervised learning methods, furthermore, the pretrained model can be used in specific downstream tasks. For video-oriented SSL methods, [24–30] show the appealing performance and potential on the task of video classification, while their positive pairs are selected from the adjacent clips within a same video. Meanwhile, most of studies are based on short videos and the quality of learned features is assessed based on video classification. Hence, it is meaningful to explore a suitable SSL scheme for tasks with long-term videos.

Video Shot Boundary Detection and Scene Segmentation. For *Video Scene Segmentation*, shot boundary detection is often conducted in advance, which is specified as a task of locating the transition positions in videos based on the similarity of the frames. 3D convolutional networks and color histogram differencing [8] are used to identify the transition boundaries. Based on the shot boundaries, [6] learns the local and global shot representations and utilizes them to split the continuous shots into scenes according to the transition of the story. More specifically, identification of each shot's segmentation point is treated as a binary classification, which is free to the location of the shot. [1] leverages unlabeled video data to obtain shot representations, which outperforms many supervised learning methods on the downstream task of *Video Scene Segmentation*. However, this method is pretrained on the entire video data of MovieNet [31] that include the testing videos, *i.e.* the

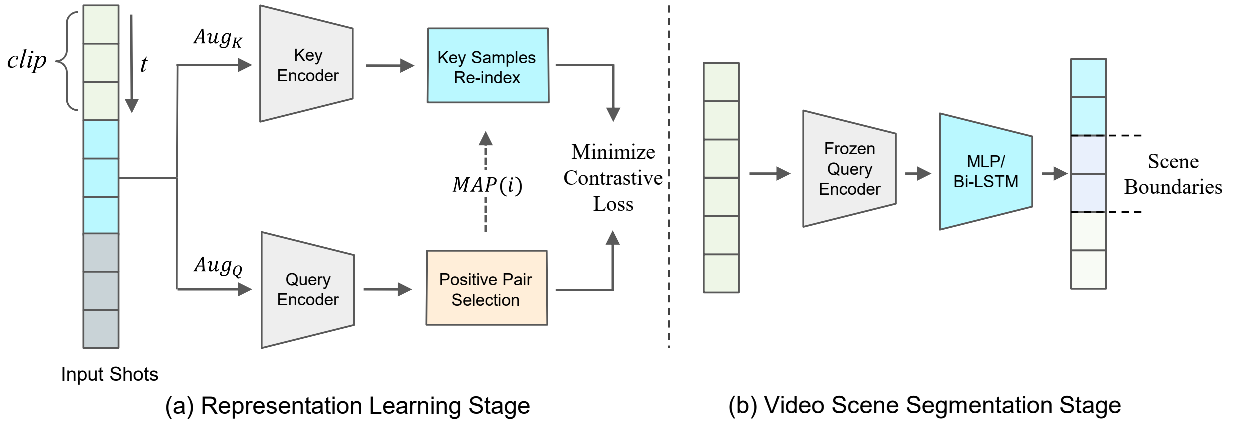

Figure 2. **The pipeline of the proposed method**. (a) Unsupervised Representation Learning Stage for learning shot representations, where $Map(i)$ is the mapping function for selecting positive samples. (b) Supervised *Video Scene Segmentation* Stage, where the quality of the shot representations is evaluated under the protocols of the non-temporal (MLP) and temporal (Bi-LSTM [19]) models.

training protocol is inconsistent with that of conventional Self-Supervised Learning methods [24] [25]. For evaluating *Video Scene Segmentation* approaches, the datasets of OVSD [32], BBC planet earth [10], MovieNet [31] and AdCuepoints [1] are frequently employed.

In this work, we propose an unsupervised representation learning method based on scene consistency and a reasonable evaluation scheme for *Video Scene Segmentation* task.

3. Methodology

As shown in Fig. 2, we aim to obtain scene consistency representations on unlabeled long-term videos and design a more reasonable benchmark to verify the quality of the extracted features on the task of *Video Scene Segmentation*. To this end, we (i) propose a Self-Supervised Learning scheme based on a novel non-temporal selection strategy to achieve scene consistency from various shots, and (ii) introduce a vanilla temporal model with less inductive bias as well as the corresponding benchmark for this segmentation task.

3.1. Consistency based Representation Learning

Approaches of Self-Supervised Learning (SSL) aim to model representation consistency to enhance network robustness against various variations, *e.g.* spatial or temporal transformations. In this work, we use an SSL framework of Siamese network to achieve the representation consistency.

More precisely, for a given query shot, the objective is to (i) maximize the similarity between the representations of query shot and positive samples, *i.e.*, key shots; (ii) minimize the similarity of the negative sample pairs if they exist. As shown in Fig. 2 (a), the input samples X are first augmented, *i.e.*, $Q = Aug_Q(X), K = Aug_K(X)$, and the i-th positive pair $\{q, k^+\}$ is formulated as follows:

$$\{q, k^+\} = \{f\left(Q[i] \mid \theta_Q\right), \; f\left(K[MAP(i)] \mid \theta_K\right)^+\} \quad (1)$$

where $[\cdot]$ stands for the indexing operation, $f\left(\cdot \mid \theta_Q\right)$ and $f\left(\cdot \mid \theta_K\right)$ are the encoders with parameters θ_Q and θ_K, respectively, $MAP(i)$ is the mapping function for selecting positive samples.

For the selection of positive samples in SSL methods based on video data, three selection strategies are frequently employed, *i.e.*, *Self-Augmented* [14], *Random* [27] and *Nearest Neighbor (NN)* [1] selections. For clarity, the three conventional selection strategies for positive samples are represented in Fig. 3 (a)-(c).

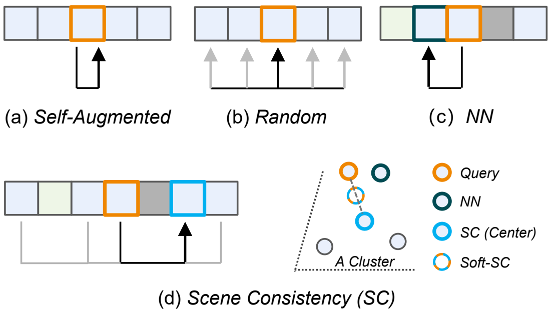

Figure 3. The illustration of four different selection strategies for positive pairs. Best viewed in color.

3.1.1 Conventional Positive Sample Selections

Self-Augmented Selection. As image-level SSL approaches, the augmented view of one shot is frequently used as its positive sample, as shown in Fig. 3 (a), the mapping function, *i.e.* the identity mapping, is employed as follows:

$$MAP_{SA}(i) = i \quad (2)$$

Random Selection. As some SSL methods [27] [24] for video classification, we select two adjacent shots of the same video as the positive pair, as shown in Fig. 3 (b), and the mapping function can be formulated as follows:

$$MAP_{RS}(i) = max(i + j, 0) \tag{3}$$

where $j \in \{-n, -n+1, \dots, n-1, n\}$ and n denotes the size of the search region around the i-th shot.

Nearest Neighborhood (*NN*) Selection. As shown in Fig. 3 (c), [1] proposed to select the positive shot with the closest representation distance to the query shot within a fixed range, and the mapping function is as follows:

$$MAP_{NN}(i) = \arg\max_{j \in I_M} f\left(Q[i] \mid \theta_Q\right) \cdot f\left(Q[max(j,0)] \mid \theta_Q\right) \tag{4}$$

where $I_M = \{i-m, \dots, i-1, i+1, \dots, i+m\}$, I_M stands for the indices of candidate samples for *NN* selection, m is the *search region size* of a given shot, and $2m + 1$ is the length of the sliding window.

3.1.2 Scene Consistency Selection

In this work, we propose the *Scene Consistency* Selection, while exploring considerable data augmentation and shuffling methods for the task of *Video Scene Segmentation*.

Positive Sample Selection with Scene Consistency. As shown in Fig. 1, for the video with non-linear narrative, previous selection methods may not work in the case that the most matching shots are far away. Therefore, we propose to select positive shot pair based on scene consistency, while the main advantage over *Random/NN* Selection is that our method is non-temporal, which is free to the shots order.

We argue that scene consistency is critical for the training on the unlabeled long-term videos due to the three reasons: (i) the similar shots in the same scene may be far away; (ii) the greater feature spacing between scenes is beneficial to the downstream task of *Video Scene Segmentation*, and it can be achieved by maximizing inter-scene distance and minimizing intra-scene distance; (iii) while the *NN* selection may result in a trivial objective, due to the maximization of the similarity of the sample pairs that maybe already the closest, the scene consistency enables the selection to achieve a more non-trivial objective.

For the proposed scene consistency-based selection, we perform online clustering of samples in a batch, and use the cluster center sample as the positive sample with respect to the query shot, as shown in Fig. 3 (d). The specified mapping function is formulated as follows:

$$MAP_{SC}(i) = \arg\min_{j \in I_C} \| f(Q[i] \mid \theta_Q) - f(Q[j] \mid \theta_Q) \|_2 \tag{5}$$

where $I_C = \{ic_1, ic_2, \dots, ic_{\#class}\}$ stands for the indices of cluster centers, $\#class$ is the number of cluster centers.

While center sample reflects the cluster-specified common information, we additionally use the query-specific individual information to generate the positive sample. Unlike the conventional multiple-instance learning [29], which treats center and query samples as multiple positive samples, we propose to construct the soft positive sample, namely *Soft-Scene Consistency (SC)* sample as follows:

$$k_{Soft-SC} = \gamma k_{SA} + (1 - \gamma) k_{SC} \tag{6}$$

where γ is a trade-off parameter, k_{SA} and k_{SC} are the key (positive) samples selected by *Self-Augmented Selection* and *Scene Consistency Selection*.

Scene Consistency Data Augmentation. Since the early stage of training is not stable, too much color augmentations, e.g. grayscale transformations, color jitter, etc., misguide the selection of positive samples, namely as *Selection Shift*. In this case, the model focuses more on non-semantic information. To solve this problem, some studies [33] directly omit color augmentations for better performance. By contrast, we propose **Asymmetric Augmentation** to alleviate the influence of *Selection Shift* and use color augmentation to further improve the performance. More specifically, augmentations without the color transformation are used in Aug_Q to get more accurate and scene consistent positive samples, while the color data augmentation operations are performed in Aug_K.

Scene Agnostic Clip-Shuffling. For fully leveraging the limited video data, we propose to construct more pseudo scene cues. In this work, the data augmentation is based on the basic unit of clip, i.e. ρ continuous shots, the generated clips are then randomly spliced disorderly for the training.

The process of *Scene Agnostic Clip Generation and Shuffling* is shown in Fig. 4.

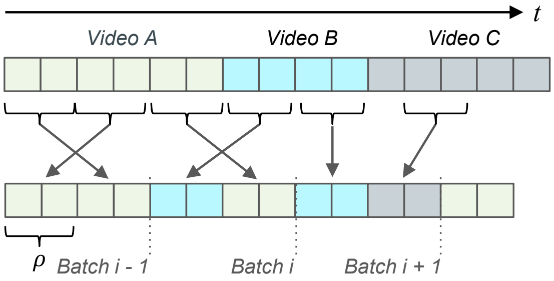

Figure 4. **The illustration of Scene Agnostic Clip-Shuffling.** Clips are spliced disorderly for training and each clip contains ρ continuous shots.

3.1.3 Negative Sample Selections

The way to choose negative samples varies according to the specific SSL frameworks. For SimCLR [12], the set of all non-positive samples within a batch is used as the negative samples, and MoCo [14] leverages a negative sample queue, which is a memory bank of previous samples output from the key encoder. However, BYOL [17] and SimSiam [16] do not use negative samples and instead resort to exploring more non-trivial solutions of SSL.

3.1.4 Objective Function

With Negative Samples. By defining $sim(\cdot,\cdot)$ as the cosine similarity, the contrastive loss function, *i.e.*, InfoNCE [11] is employed and formulated as follows:

$$L_{con} = -\log \frac{\sum_{k\in\{k^+\}} e^{(\mathrm{sim}(q,k)/\tau)}}{\sum_{k\in\{k^+,k^-\}} e^{(\mathrm{sim}(q,k)/\tau)}} \tag{7}$$

where k^+ and k^- stand for the positive and negative samples for the query q, and the τ is the temperature term [34].

Without Negative Samples. By maximizing the similarity between the query and positive samples, the contrastive loss without negative samples is formulated as follows:

$$L_{con} = -2 \sum_{k\in\{k^+\}} (\mathrm{sim}(\mathcal{P}_\theta(q), k_{SG}) + \mathrm{sim}(\mathcal{P}_\theta(k), q_{SG})) \tag{8}$$

where $\mathcal{P}_\theta$ is the predictor $\mathcal{P}$ with parameters θ [16, 17], k_{SG} and q_{SG} are the samples with stop-gradient (SG) [16, 17].

3.2. Video Scene Segmentation

After the unsupervised pretraining, two downstream models are used to evaluate the quality of the extracted features with the frozen query encoder.

Problem Definition. For the *Video Scene Segmentation*, [6] [1] convert the task into a binary classification task of shot semantics by modeling the temporal relationship of adjacent shot features. In this way, we can determine whether the end of a shot is the end of a scene story.

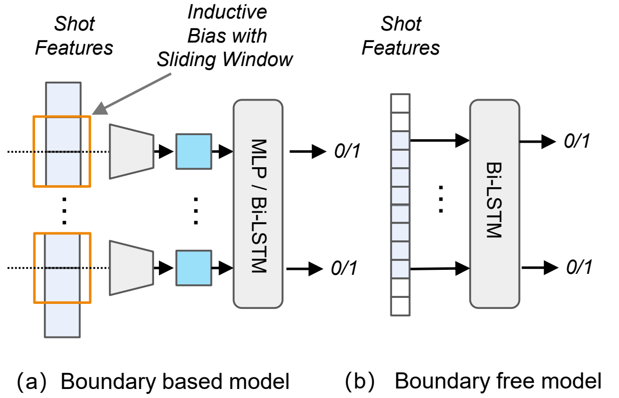

Figure 5. The illustration of boundary based model (a) and boundary free model (b) for *Video Scene Segmentation*.

Boundary free model. While the previous downstream task of *Video Scene Segmentation* is concluded to a shot boundary modeling based approach, as shown in Fig. 5 (a), we introduce a vanilla boundary-free model. As shown in Fig. 5 (b), the proposed model covers the long-term dependence of shot representations based on sequence-to-sequence learning. Compared with boundary based model in Fig. 5 (a) that introduces inductive bias for the shot boundary modeling with the sliding windows, the suggested model (b) takes the shot features as the basic temporal input unit, enabling the model to explore both local and global semantic relations.

4. Experiments

4.1. Experimental Setup

Dataset. MovieNet [31] consists of 1,100 movies with a large amount of multi-modal data and annotations, and the total duration of all movies is about 3000 hours, it is the largest dataset for movie understanding analysis by far. Besides, MovieNet [31] is split into a training set with 660 movies, a validation set with 220 movies and a testing set with 220 movies. Currently, for the task of *Video Scene Segmentation*, 190, 64 and 64 videos are labeled with scene boundaries for the training, validation and test sets, respectively. More importantly, ***movies in the MovieScene [6] are all included in MovieNet [31].***

It is worth noting that there are two versions of annotation about *Video Scene Segmentation* task associated with MovieNet, one with only 150 annotations (called *MovieScenes* in [1, 6], used in earlier methods [6] but it is no longer available), and one with a total of 318 annotations (abbreviated as *MovieScenes-318* in this work). Since the small scale of of BBC [10] and OVSD [32] datasets and unavailability of AdCuepoints [1] dataset, we instead adopt MovieNet [31] dataset to evaluate the related approaches, more details are in the *Supplementary Materials*.

Representation Learning Stage. For visual modality, each shot consists of 3 keyframes and ResNet50 [38] is chosen as the default backbone to learn the shot representations. The audio backbone used in [6] is applied for audio modality, more details about the backbone encoders can be found in *Supplementary Materials*.

For pretraining data, (i) training set (660 movies) in MovieNet [31] is used to learn the shot representations, while we also conduct experiments with (ii) all data (1,100 movies) [1] for a fair comparison. In particular, although test data without the scene boundary labels are used for representation learning in setting (ii), **it is not recommended** to use all the data for pretraining because we usually have no prior access to test data in real scenarios. Moreover, for the most of Self-Supervised benchmarks [11–18], representation learning is performed only on the training set, rather than all of the data.

Video Scene Segmentation Stage. For existing Self-Supervised methods on images and videos, a simple downstream model is frequently used to evaluate the representation quality of the frozen encoders. For instance, a linear

Table 1. Results of supervised methods w/o SSL for the task of Video Scene Segmentation on MovieNet.

Methods	Dataset	AP	F1
SCSA [9]	M.S.	14.7	-
Story Graph [35]	M.S.	25.1	-
Siamese [10]	M.S.	28.1	-
ImageNet [36]	M.S.	41.26	-
Places [37]	M.S.	43.23	-
LGSS [6]	M.S.	47.1	-
LGSS w/o DP [6]	M.S.	44.9	-
LGSS w/o DP [6] *	M.S-318	44.9	38.52

* Our implementations based on official public codebase,[2] while *DP (Dynamic Programming)* isn't public available.

Table 2. Results of methods w/ SSL for the task of Video Scene Segmentation on MovieNet.

Methods	Pretrain Data	Eval.	Protocol	AP	F1
ShotCoL [1]	Train.+Test.+Val.	M.S	MLP [1]	52.83	-
ShotCoL [1]	Train.+Test.+Val.	M.S-318	MLP [1]	53.37	-
ShotCoL [1] *	Train.+Test.+Val.	M.S-318	MLP [1]	52.89	49.17
SCRL (ours)	Train.+Test.+Val.	M.S-318	MLP [1]	54.82	51.43
ShotCoL [1] *	Train. only	M.S-318	MLP [1]	46.77	45.78
SCRL (ours)	Train. only	M.S-318	MLP [1]	53.74	50.40
ShotCoL [1] *	Train. only	M.S-318	Bi-LSTM	48.21	46.52
SCRL (ours)	Train. only	M.S-318	Bi-LSTM	**54.55**	**51.39**

* Our implementations.

fully-connected layer is widely used for evaluation. However, for the *Video Scene Segmentation* task, we cannot determine whether the ending position of a single shot is the scene boundary or not. Consequently, a boundary-based non-temporal model (MLP-based protocol, followed by [1]) and a boundary-free temporal model (Bi-LSTM [19]-based protocol, proposed by us) are employed to evaluate the capability of the encoder for local-to-global modeling.

Metrics. We use the mean of Average Precision (AP) [6] [1] specified to ground truth scene boundaries of each movie, as well as F1-score for the evaluation.

Implementation Details. During the learning stage of Self-Supervised representation, the batch size is set to 1,024 (shots), initial learning rate is set to 0.03 and the training epoch is 100. The parameters of the visual and audio encoders are randomly initialized. Besides, we perform naive *K-Means* algorithm [39, 40] for online clustering and the cluster number $\#class$ is set to 24, while the clip length, *i.e.* ρ of Scene Agnostic Clip Shuffling is set to 16. MoCov2 [14] with the queue size of 65,536, momentum value of 0.999, temperature of 0.07 and cosine learning rate decay, are used as our SSL framework setting. For the *Video Scene Segmentation* task, $num\text{-}of\text{-}shot$ [1] is set to 4 and 40 for the MLP [1] and Bi-LSTM protocols, respectively. Each pretraining trial is conducted on the server with 8 NVIDIA V100 GPUs for approximate 24 hours in visual modality and 10 hours in audio modality. The dimensions of visual and audio features used for both pretraining and evaluation are 2,048 and 512, respectively. More details, *e.g.*, the choice of hyperparameter, are presented in *Supplementary Materials*.

4.2. Comparison with Existing Methods

Tables 1 and 2 present an overall performance of methods w/ or w/o SSL for the *Video Scene Segmentation* task, where *M.S.* stands for *MovieScenes* dataset with 150 annotated movies, and *Eval.* means the dataset used for supervised *evaluation stage* after the pretraining. Besides, *Train., Test., and Val.* represent *training, testing* and *validation* sets of MovieNet [31].

We have reproduced the performance of ShotCoL [1] on the entire dataset (1,100 movies) for comparison, although it is suggested to conduct the pretraining stage only on the training set. Compared with ShotCoL [1] that has a decline of 6.12 in terms of AP, our method can achieve competitive performance with less training data, with only a decline of 1.08 in terms of AP. The proposed method outperforms the supervised state-of-the-art method, *i.e.,* LGSS [6] by margins of 9.65 in terms of AP and 12.87 in terms of F1.

4.3. Ablation Study

We perform all the ablation experiments using only the training data of MovieNet in SSL stage, and evaluate the performance on downstream task based on MLP protocol for fairness.

Positive Sample Selection. We first conduct ablation experiments on the four different selection methods of positive pairs, *i.e.*, *Self-Augmented Selection*, *Random Selection*, *Nearest Neighborhood (NN) Selection* and *Scene Consistency (SC) Selection*. Tab. 3 shows that *Scene Consistency Selection* method achieves better performance than the other selection methods, which outperforms the state-of-the-art algorithm [1] by a margin of 2.95 in terms of AP. Meanwhile, the loss evolution curves of above methods are shown in Fig. 6. We can find that *Self-Augmented Selection* reaches the lowest loss value, while obtaining the worst performance on the task of *Video Scene Segmentation*. Due to the trivial objective introduced by *NN Selection* that is discussed in Section 3.1.2, it achieves the fastest convergence rate during the early training, while stagnating to a mediocre performance. By contrast, *SC Selection* has a relatively moderate convergence rate, and achieves the best performance among all the selection strategies.

[2] https://github.com/AnyiRao/SceneSeg

Table 3. Ablation results of Positive Sample Selection.

Methods	Selection Strategy	AP
MoCo [14]	Self-Augmented	42.51
-	Random ($n = 1$)	43.24
ShotCol [1]	NN ($m = 8$)	46.77
SC	Scene Consistency	**49.71**

Table 4. Ablation results of SSL methods w/ and w/o Scene Agnostic Clip-Shuffling.

Methods	w/o	w/	AP
NN	✓	×	46.77
NN	×	✓	48.63
SC	✓	×	49.71
SC	×	✓	**52.17**

Table 5. Ablation results of Multiple Positive Samples (MPS).

Methods	Positive Sample(s)	AP
SC	Center	52.17
MPS-SC	Self and Center	51.20
Soft-SC	Eq. 6 ($\gamma = 0.5$)	**53.74**

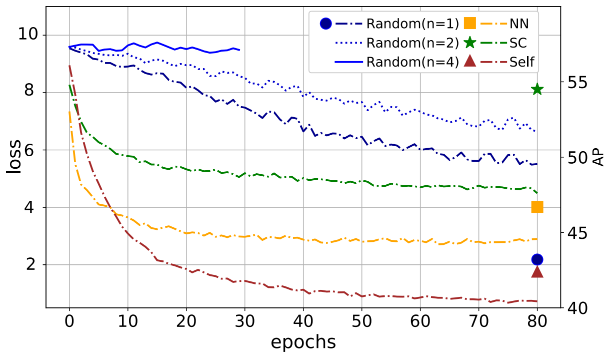

Figure 6. **Loss evolution curves and AP results** of the training with different selection strategies.

Scene Agnostic Clip-Shuffling. The results of ablation study specific to the Scene Agnostic Clip-Shuffling are presented in Tab. 4. Tab. 4 shows that the proposed *Clip-Shuffling* achieves improvements of 1.85 and 2.46 in terms of AP for *NN* and *SC* methods, respectively. These results verify the advantage of the proposed positive sample selection discussed in Section 3.1.2 that *SC* is free to the shot order in a video.

Multiple Positive Samples (MPS). Moreover, we study the performance of multiple positive samples in Tab. 5. As shown in Tab. 5, *Soft-SC* achieves the best performance of 53.74 in terms of AP. Although single positive sample is employed in SC, it still achieves better performance than MPS-SC that employ multiple positive samples.

4.4. Analysis of the Proposed Method

Generalizability to the large-scale supervised approach. To study the generalizability of the proposed method, we equip our trained models with LGSS [6], where LGSS is a large-scale supervised method and utilizes various pretrained models with multi-modality. As is Shown in Tab. 6, our trained model, trained only on the unlabeled training set, and based on the same backbone, *i.e.* ResNet-50 [38], achieves an improvement of 4.0 in terms of AP over the approach without our trained model.

Table 6. Generalizability to the large-scale supervised approach.

Methods	Modalities	AP
LGSS [6] w/o SSL	**Visual(Place, ResNet50)** +Action+Actor+Audio	44.9
LGSS [6] w/ SSL	**Visual(SSL, ResNet50)** +Action+Actor+Audio	**48.9**

Performance on different Self-Supervised Learning (SSL) frameworks. Four popular SSL frameworks are used for evaluating our method, *i.e.*, SimCLR [12], MoCo [14], BYOL [17] and SimSiam [16]. Tab. 7 shows that the SSL framework with momentum updates and negative samples achieves the best performance for the *Video Scene Segmentation* task. Due to the momentum update mechanism, the proposed method embedded in the framework of BYOL [17] achieves an improvement of 10.53 over that in SimSiam [16], and a similar conclusion is reached in [27].

Table 7. Results of the proposed method based on various Self-Supervised Frameworks.

Methods	SSL Frameworks	w/ negative samples	w/ momentum update	AP
SCRL	SimSiam [16]	×	×	39.82
SCRL	SimCLR [12]	✓	×	45.32
SCRL	BYOL [17]	×	✓	50.35
ShotCoL [1]	MoCo [14]	✓	✓	46.77
SCRL	MoCo [14]	✓	✓	**53.74**

Boundary free model for evaluation. To study the performance of the introduced boundary free model, the proposed method under MLP and Bi-LSTM protocols for the scene segmentation task is evaluated in Tab. 8. Since Bi-LSTM protocol has less inductive bias than *sliding window* based MLP protocol, it is able to model representations of longer shot, hence achieves better performance on the task of *Video Scene Segmentation*. More specifically, the performance of Bi-LSTM protocol increases as the length of the shots increases, while the performance of MLP protocol decreases instead. More details can be found in *Supplementary Materials*.

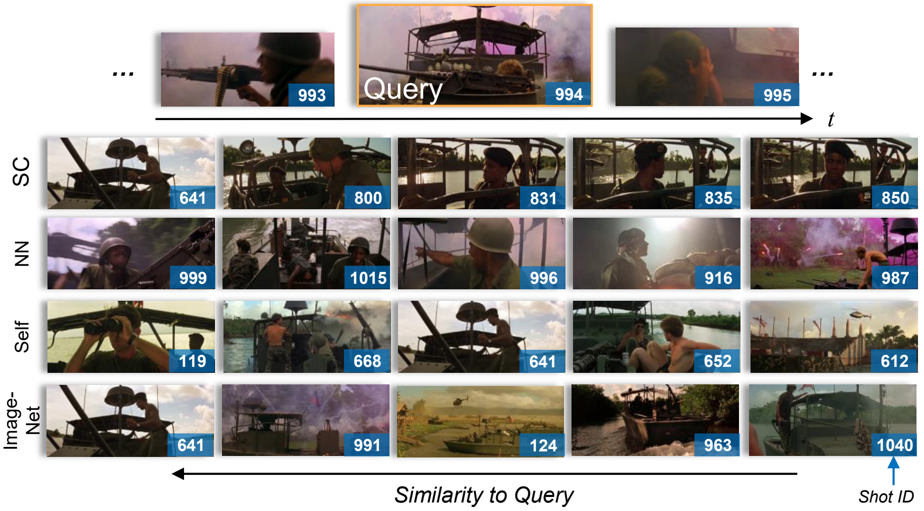

Figure 7. **The visualization results of shot retrieval**. Overall, ***NN*** tends to select adjacent shots, ***Self*** shows less relevance to the query and ***ImageNet*** retrieves many kinds of boats. Compared with the other methods, the results of ***SC*** are more consistent in the semantic information, *i.e.*, there is a man staying in the boat, and ***SC*** achieves a larger span (*i.e.*, from 641 to 850) than ***NN*** according to shot IDs. Meanwhile, ***SC*** shows better robustness against the interference of pink smoke in the 994-th shot as the results are more pure.

Table 8. Results of *Video Scene Segmentation* using the proposed method under MLP and Bi-LSTM protocols.

Protocols	Shot-Len	AP	F1	#Param
MLP [1]	4	53.74	50.40	37.75 M
MLP [1]	10	49.61 ↓	44.04 ↓	88.09 M ↑
Bi-LSTM	10	43.94	42.12	15.22 M
Bi-LSTM	40	**54.55** ↑	**51.39** ↑	15.22 M

Visualization of Shot Retrieval. To get more intuition for the proposed selection, we conduct retrieval experiments using four selection methods, *i.e.*, *SC, NN, Self-Augmented and ImageNet selections*, and present the results in Fig. 7. More specifically, for a given shot, we calculate the similarities between it and the other shots in the entire movie, then visualize the TOP-5 most similar shots in Fig. 7.

4.5. Limitations

Multi-modal Pretraining. In order to test the performance of the proposed algorithm generalizing to multi-modal data, we also conduct experiments with audio and visual modalities in the SSL stage, the joint multi-modal learning scheme follows [41]. However, we did not achieve any improvement and were confronted with the same concern that is mentioned in [1], as shown in Table. 9. Possible reasons are that (i) the publicly available audio data of each shot are incomplete, (ii) the raw audio data are not available yet due to copyright restrictions [31] and (iii) LGSS [6] utilizes various pretrained models on the other datasets, while the methods in the comparison are trained from scratch. Therefore, it is meaningful to shed light on how to pretrain better multi-modal representations on the MovieNet [31].

Table 9. AP results of the multi-modal experiment on MovieNet. Backbones of following methods for each modality are the same.

Methods	Visual	Audio	Visual+Audio
LGSS [6]	39.0	17.5	43.4
ShotCoL [1]	46.77	27.92	44.32
SCRL	53.74	29.39	50.80

5. Conclusion

We present a Self-Supervised Learning (SSL) scheme based on Scene Consistency to obtain better shot representations for the unlabeled long-term videos. The proposed method achieves the state-of-the-art performance on the task of *Video Scene Segmentation* under various protocols, and significant better generalization performance when it is equipped with large-scale supervised approach. Besides, we introduce a fair pretraining protocol and a more comprehensive evaluation metric for the task of *Video Scene Segmentation*, to make the assessment of the SSL more meaningful in practice.

Acknowledgment The work was supported by the National Natural Science Foundation of China under grants no. 61602315, 91959108, the Science and Technology Project of Guangdong Province under grant no. 2020A1515010707, the Science and Technology Innovation Commission of Shenzhen under grant no. JCYJ20190808165203670.

References

[1] Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. Shot contrastive self-supervised learning for scene boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9796–9805, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[2] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 2

[3] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13526–13535, 2021. 2

[4] Anthony Cioppa, Adrien Deliege, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B Moeslund. A context-aware loss function for action spotting in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13126–13136, 2020. 2

[5] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1711–1721, 2018. 2

[6] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10155, 2020. 2, 5, 6, 7, 8

[7] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, Miguel Bugalho, and Isabel Trancoso. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(8):1163–1177, 2011. 2

[8] Jakub Lokoč, Gregor Kovalčik, Tomáš Souček, Jaroslav Moravec, and Přemysl Čech. A framework for effective known-item search in video. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1777–1785, 2019. 2

[9] Vasileios T Chasanis, Aristidis C Likas, and Nikolaos P Galatsanos. Scene detection in videos using shot clustering and sequence alignment. *IEEE transactions on multimedia*, 11(1):89–100, 2008. 2, 6

[10] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. A deep siamese network for scene detection in broadcast videos. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1199–1202, 2015. 2, 3, 5, 6

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2, 5

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 4, 5, 7

[13] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 33:22243–22255, 2020. 2, 5

[14] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 3, 4, 5, 6, 7

[15] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 2, 5

[16] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2, 4, 5, 7

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 2, 4, 5, 7

[18] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2, 5

[19] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015. 3, 6

[20] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR 2018*, 2018. 2

[21] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[22] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[23] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2

[24] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 2, 3

[25] Ali Diba, Vivek Sharma, Reza Safdari, Dariush Lotfi, Saquib Sarfraz, Rainer Stiefelhagen, and Luc Van Gool. Vi2clr: Video and image for visual contrastive learning of representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1502–1512, 2021. 2, 3

[26] Haofei Kuang, Yi Zhu, Zhi Zhang, Xinyu Li, Joseph Tighe, Soren Schwertfeger, Cyrill Stachniss, and Mu Li. Video contrastive learning with global context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3195–3204, 2021. 2

[27] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021. 2, 3, 7

[28] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[29] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 2, 4

[30] Adria Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altché, Michal Valko, et al. Broaden your views for self-supervised video learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1255–1265, 2021. 2

[31] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer, 2020. 2, 3, 5, 6, 8

[32] Daniel Rotman, Dror Porat, and Gal Ashour. Optimal sequential grouping for robust video scene detection using multiple modalities. *International Journal of Semantic Computing*, 11(02):193–208, 2017. 3, 5

[33] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10075–10085, 2021. 4

[34] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 5

[35] Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. Storygraphs: visualizing character interactions as a timeline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 827–834, 2014. 6

[36] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[37] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 6

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 7

[39] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. 6

[40] Haozhe Liu, Haoqian Wu, Weicheng Xie, Feng Liu, and Linlin Shen. Group-wise inhibition based feature regularization for robust classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 478–486, 2021. 6

[41] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2(6):7, 2020. 8