Towards Global Video Scene Segmentation with Context-Aware Transformer

Yang Yang[1,2,3*], Yurui Huang[1†], Weili Guo[1], Baohua Xu[4], Dingyin Xia[4]

[1]Nanjing University of Science and Technology
[2]MIIT Key Lab. of Pattern Analysis and Machine Intelligence, NUAA
[3]State Key Lab. for Novel Software Technology, NJU
[4]HUAWEI CBG Edu AI Lab
{yyang, huangyurui, wlguo}@njust.edu.cn, {xubaohua1, xiadingyin}@huawei.com

Abstract

Videos such as movies or TV episodes usually need to divide the long storyline into cohesive units, i.e., scenes, to facilitate the understanding of video semantics. The key challenge lies in finding the boundaries of scenes by comprehensively considering the complex temporal structure and semantic information. To this end, we introduce a novel Context-Aware Transformer (CAT) with a self-supervised learning framework to learn high-quality shot representations, for generating well-bounded scenes. More specifically, we design the CAT with local-global self-attentions, which can effectively consider both the long-term and short-term context to improve the shot encoding. For training the CAT, we adopt the self-supervised learning schema. Firstly, we leverage shot-to-scene level pretext tasks to facilitate the pre-training with pseudo boundary, which guides CAT to learn the discriminative shot representations that maximize intra-scene similarity and inter-scene discrimination in an unsupervised manner. Then, we transfer contextual representations for fine-tuning the CAT with supervised data, which encourages CAT to accurately detect the boundary for scene segmentation. As a result, CAT is able to learn the context-aware shot representations and provides global guidance for scene segmentation. Our empirical analyses show that CAT can achieve state-of-the-art performance when conducting the scene segmentation task on the MovieNet dataset, e.g., offering 2.15 improvements on AP.

Introduction

With the development of internet, a significantly increasing number of videos have been produced and stored. In order to reduce manual costs and improve efficiency, intelligent video understanding has received extensive attention and researches (Yang et al. 2018; Zhu et al. 2020). A fundamental aspect of video semantic understanding is scene segmentation (i.e., scene boundary detection) (Rao et al. 2020; Chen et al. 2021; Huang et al. 2020; Wang et al. 2021), which plays an important role in facilitating downstream tasks. For example, students can quickly locate knowledge points according to the segmented educational videos; users can utilize the interested scenes to retrieve movies with similar themes; video platforms can advertise based on the segmentation point to obtain higher revenue. Compared with locating a shot directly using visual cues (Cotsaces, Nikolaidis, and Pitas 2006) (here shot represents a set of visually continuous frames over an uninterrupted period of time), scene segmentation is a more challenging task, that aims to find the temporal locations of scene with complex temporal structure and semantic information (here scene denotes a sequence of shots to describe a semantically associated story).

Despite the great progress in temporal localization, most existing approaches usually focus on localizing certain action from short videos (Lin et al. 2019, 2018; Long et al. 2020), these methods usually pre-define a list of categories that are visually distinguishable (Rao et al. 2020). However, scene segmentation poses significantly more difficult challenges: 1) Coarse-grained labels. The input video has only binary boundary labels, without the fine-grained content categories for each scene as action recognition. 2) High-order coherence. Scene segmentation needs to group the shots by considering extracted high-order information, i.e., semantical coherence, rather than simple visual continuity. To address these challenges, unsupervised approaches (Baraldi, Grana, and Cucchiara 2015; Chasanis, Likas, and Galatsanos 2009) were firstly developed, which detected the boundary with pairwise similarity comparison or nearest shots clustering. Nevertheless, their performance is relatively low considering the unsupervised setting. Furthermore, (Rao et al. 2020; Chen et al. 2021; Das and Das 2020) introduced the scene boundary label for supervised prediction with contextual shots within the sliding window. However, these methods are limited to using labeled data, without considering the unlabeled videos. Therefore, self-supervised approaches (Chen et al. 2020; Roh et al. 2021) are widely researched by learning effectiveness representation without relying on costly ground-truth annotations. Therefore, the self-supervised learning methods (Chen et al. 2021; Mun et al. 2022; Wu et al. 2022a) have been designed to employ the pre-training protocols for learning spatio-temporal patterns in video scenes. However, in current self-supervised methods, the strategy of pretext task designs and high-level contextual representation modeling are not well addressed.

Therefore, in this paper, we develop a Context-Aware Transformer (CAT), which takes advantage of the princi-

[*]corresponding authors.
[†]work is done during the internship at HUAWEI CBG Edu AI Lab

ple behind video production that nearby shots should have semantically cohesive story-arch, and the far away shots will have a transition with little similarity. In detail, CAT develops the local-global self-attention heads to synthesize the complementary information from both long-term and short-term neighbors, rather than encoding the shots with only single-level contextual information. Moreover, in self-supervised training CAT, we propose shot-to-scene level pretext tasks, i.e., Shot Masking Prediction, Shot Order Prediction, Global Scene Matching, and Local Scene Matching, that leverage pseudo-boundaries to capture semantic contextual representation during pre-training, thus leading to precise scene boundary detection in fine-tuning stage with labeled data. Along this line, we can overcome the limitation of modeling videos by learning context-aware shot representations, and wisely employing the unlabeled videos. Consequently, CAT provides global guidance for video scene segmentation.

Related Work

Video Scene Segmentation (also known as scene boundary detection) aims at identifying the begin and end locations of different scenes with cohesive story-arch in videos. Early attempts mainly adopt the unsupervised learning to contrast or cluster neighboring shots into scenes. For example, (Rui, Huang, and Mehrotra 1998) grouped shots into semantically related scenes with time-adaptive similarity. (Rasheed and Shah 2003) utilized the motion content, shot length and color properties of shots for first pass cluster, then computed scene dynamics for fine-grained cluster. (Chasanis, Likas, and Galatsanos 2009) proposed an improved spectral clustering method and employed the fast global k-means algorithm for grouping shots. (Baraldi, Grana, and Cucchiara 2015) introduced a deep siamese network for segmenting videos into coherent scenes. (Sidiropoulos et al. 2011; Yang et al. 2021; Baraldi, Grana, and Cucchiara 2015) fused multi-modal such as audio and visual features for final detection. However, these methods always rely on manually designed similarity mechanisms, which are suffered from low performance and efficiency. Therefore, supervised approaches are researched, which adopted the boundary label for supervised training. For example, (Rotman, Porat, and Ashour 2017) formulated the scene detection as a generic optimization problem to optimally group shots into scenes. (Das and Das 2020) concatenated a shot with its left and right contexts for segment boundary prediction. (Rao et al. 2020) proposed to hierarchically learn shot embedding to provide a top-down scene segmentation with multi-modal information. Furthermore, to utilize the unlabeled videos, self-supervised segmentation approaches are proposed. For example, (Chen et al. 2021) presented a self-supervised shot embedding approach to learn a shot representation that maximizes the similarity between nearby shots compared to randomly selected shots. (Mun et al. 2022) pre-trained a transformer encoder with pseudo-boundaries, and then fine-tuned the encoder with labeled data. Nevertheless, these methods always adopted sophisticated model architectures, without carefully considering the contextual information of the long-term video.

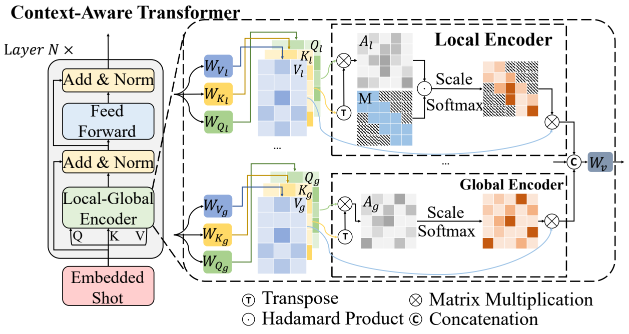

Figure 1: Illustration of Context-Aware Transformer (CAT). CAT designs the local-global self-attentions to comprehensively consider both short-term shots and potential long-term correlated shot information.

Self-Supervised Representation Learning attempts to learn representations using unlabeled data by solving pretext tasks using pseudo-supervised learning. The pseudo-labels are automatically created without requiring labeled data (Jing and Tian 2021). For example, (Pathak et al. 2016; Vincent et al. 2008) used the pretext tasks of reconstructing corrupted, (Doersch, Gupta, and Efros 2015) proposed to classify inputs with pseudo-labels. Inspired by self-supervised learning, many approaches are proposed with various pretext tasks in video understanding tasks. For example, (Ahsan, Sun, and Essa 2018) presented a pretext task of masked frame modeling to learn temporal dependency between frames. (Xu et al. 2019) discovered the spatiotemporal representations of the video by predicting the order of shuffled clips from the video. (Kuang et al. 2021) proposed a video-level contrastive learning method based on segments to formulate positive pairs. However, most methods concentrated on the classification task by modeling the shot level pretext tasks, which may be sub-optimal to the video scene segmentation task.

Proposed Method

Considering that state-of-the-art methods (Rao et al. 2020; Chen et al. 2021; Mun et al. 2022) formulate scene segmentation based on the constituent set of shots (i.e., determine whether a shot boundary is a scene boundary), all input videos are first divided into shots with standard shot detection techniques (Sidiropoulos et al. 2011), details are in the supplementary. Therefore, given an untrimmed video $\{\mathbf{v}_t\}_{t=1}^{T}$ with T shots, where $\mathbf{v}_t$ is the $t-$th shot. Scene segmentation is to generate a set of boundary label $\mathbf{y} = \{y_t\}_{t=1}^{T}$, where $y_t = 1$ represents the boundary, otherwise $y_t = 0$. To this end, we first introduce a context-aware Transformer encoder to model the contextual information, and then propose a self-supervised learning scheme with shot-to-scene pretext tasks to learn discriminative shot representations for segmentation.

Context-Aware Transformer

The key challenge to encoding the sequential shots is that: different shot neighbors have various importance in modeling contextual information. To overcome this challenge, we design the Transformer with local-global self-attention heads to integrate both short-term shots and potential long-term correlated shot information.

Shot Encoder. Following (Rao et al. 2020; Chen et al. 2021; Wu et al. 2022b; Mun et al. 2022), we employ a shot encoder f_e to encode a shot by capturing its spatio-temporal patterns. Given a shot $\mathbf{v}_t$, the representations can be formulated as: $f_e(\mathbf{v}_t)$. Then the encoded shot sequence is sent into the CAT.

Local Encoder. To comprehensively encode each shot by considering the dependencies between shots, we employ the transformer encoder (Vaswani et al. 2017) as the backbone, which can encode the relationships among independent shots by adopting the self-attention mechanism. Specifically, as shown in Figure 1, with the input shots, a video can be denoted as $\bar{\mathbf{v}} = [\bar{\mathbf{v}}_1, \bar{\mathbf{v}}_2, \cdots, \bar{\mathbf{v}}_T] = f_e(\mathbf{v})\bar{W} \in \mathcal{R}^{T \times d}$, where d is the hidden dimension, $\bar{W} \in \mathcal{R}^{d_1 \times d}$ is the learnable matrix. In the self-attention layer, the input representations can be used to compute three matrices: Q, K, and V corresponding to queries, keys, and values. Note that local self-attention heads only calculate the dot-product similarities between queries and keys of the shot neighbors, i.e., a fixed L-size window centered on the shot:

$$\begin{aligned} &Q_l = \bar{\mathbf{v}} W_{Q_l}, \quad K_l = \bar{\mathbf{v}} W_{K_l}, \quad V_l = \bar{\mathbf{v}} W_{V_l}, \\ &A_l = \frac{Q_l K_l^\top}{\sqrt{d_{N_l}}} \cdot M \quad Att(\bar{\mathbf{v}}_l) = \sigma(A_l) V_l, \end{aligned} \tag{1}$$

where $Q_l \in \mathcal{R}^{T \times d_{N_l}}$, $K_l \in \mathcal{R}^{T \times d_{N_l}}$, $V_l \in \mathcal{R}^{T \times d_{N_l}}$, and $W_{Q_l} \in \mathcal{R}^{d \times d_{N_l}}, W_{K_l} \in \mathcal{R}^{d \times d_{N_l}}, W_{V_l} \in \mathcal{R}^{d \times d_{N_l}}$ are learnable matrices. N_l denotes the number of local heads. The activation function σ can be used as softmax here. $M \in \mathcal{R}^{T \times T}$ represents the mask matrix with padding, where the shots in defined local window is 1, otherwise is 0.

Global Encoder. To introduce the extra neighbor shot as complementary information, we further propose jointly modeling strategy with global self-attention heads, which directly determine attention distributions with the dot-product similarity between fully queries and keys:

$$\begin{aligned} &Q_g = \bar{\mathbf{v}} W_{Q_g}, \quad K_g = \bar{\mathbf{v}} W_{K_g}, \quad V_g = \bar{\mathbf{v}} W_{V_g}, \\ &A_g = \frac{Q_g K_g^\top}{\sqrt{d_{N_g}}} \quad Att(\bar{\mathbf{v}}_g) = \sigma(A_g) V_g, \end{aligned} \tag{2}$$

where $Q_g \in \mathcal{R}^{T \times d_{N_g}}$, $K_g \in \mathcal{R}^{T \times d_{N_g}}$, $V_g \in \mathcal{R}^{T \times d_{N_g}}$, and $W_{Q_g} \in \mathcal{R}^{d \times d_{N_g}}, W_{K_g} \in \mathcal{R}^{d \times d_{N_g}}, W_{V_g} \in \mathcal{R}^{d \times d_{N_g}}$ are learnable matrices. N_g denotes the number of global heads. Finally, local-global self-attention is composed of $N = N_l + N_g$ parallel heads, and $d_{N_l} = d_{N_g} = d/N$.

In summary, the local head attentions are responsible for capturing local dependencies based on local details, i.e., the potential intra-scene shots, and global head attentions are designed to model the long-term dependencies between shots, i.e., the potentially missed intra-scene and inter-scene shots. The combination of local and global attention enables our CAT to dynamically model local shots and capture the global dependencies of similar shots. Consequently, we can acquire output representations of shots, i.e., $\hat{\mathbf{v}} = f_t(\bar{\mathbf{v}}) = [\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \cdots, \hat{\mathbf{v}}_T] \in \mathcal{R}^{T \times d}$, where f_t denotes the CAT.

Self-Supervised Training

To employ unlabeled videos, there are two ways: semi-supervised and self-supervised techniques. A direct way in semi-supervised methods (Arazo et al. 2020) is to use a model's predictions to obtain artificial labels for unlabeled data. A specific variant is the pseudo labeling, which converts the model predictions of unlabeled data to hard labels for calculating the cross entropy. However, when we used semi-supervised ideas in the early stage, we found that the model performance was not as good as the supervision effect. We find that the reason is that the distributions of unlabeled and labeled data are inconsistent in the MovieNet dataset, which leads to the problem of noisy labeling predicted by the model trained on labeled data. In detail, we set the supervised data with label 1, and the unsupervised data with label 0, then we train a binary classifier using the representations $f_e(\mathbf{v})$. The result of test AUC is 78.6, which indicates that we can easily distinguish the supervised and unsupervised data, i.e. existing distribution drift problem. Therefore, the self-supervised scheme, which first trains a generalized model using the unsupervised data and then finetunes the pre-trained model using supervised model, can well overcome this challenge.

Pre-Training Objectives

Shot Masked Modeling (SMM). Inspired by masked language modeling (Vaswani et al. 2017), we adopt the shot masked modeling task that reconstructs the representation of masked shots based on the surrounding shots. In detail, given the input shot features $\bar{\mathbf{v}}$, we randomly mask them with a probability of 15%. For masked shot sets, we learn to reconstruct the output representations to their input shot features with a regression head. The reconstruction loss can be formulated as:

$$L_{smm} = \sum_{i \in \mathcal{D}_m} \|\bar{\mathbf{v}}_i - h_{smm}(\hat{\mathbf{v}}_i)\|_2^2 \tag{3}$$

where h_{smm} is a regression head to match the contextualized shot representations with input features $\bar{\mathbf{v}}_i$. $\hat{\mathbf{v}}_i$ denotes the learned representation of i-th shot by CAT. $\mathcal{D}_m$ denotes the set of masked shots.

Shot Order Modeling (SOM). SOM aims at full-scale exploiting the sequential nature of video input. Inspired by the frame order modeling (Li et al. 2020), we randomly select 15% of the shots to be shuffled, and the SOM is to reconstruct their original timestamps, i.e., $\mathbf{s} = \{s_j\}_{j=1}^{|\mathcal{D}_o|}$, where $s_j \in \{1, 2, \cdots, T\}$, $\mathcal{D}_o$ is the set of shuffled shots and $|\mathcal{D}_o|$ represents the size. SOM can be formulated as a classification problem, where $\mathbf{s}$ is the ground-truth labels of the reordered shots. The objective can be formulated as:

$$L_{som} = \sum_{j \in \mathcal{D}_o} CE(s_j, h_{som}(\hat{\mathbf{v}}_j)) \tag{4}$$

where CE represents the cross-entropy loss, h_{som} denotes the order predictor with a softmax layer.

Global Scene Matching (GSM). GSM aims to make the shot representations similar to its associated scene, while

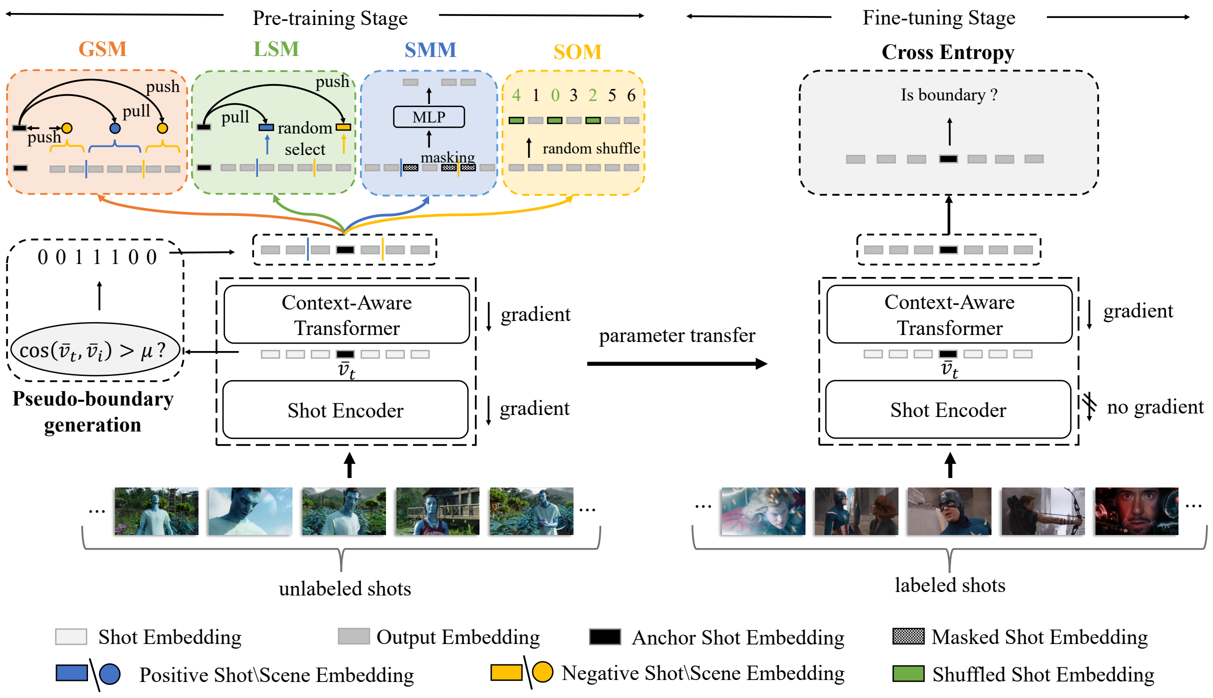

Figure 2: Illustration of self-supervised training. In the pre-training stage, we train the shot encoder and context-aware Transformer with shot-to-scene pretext tasks using pseudo boundary in an unsupervised manner. Then, we fix the shot encoder, and fine-tune the context-aware Transformer with supervised data, including boundary prediction and supervised contrastive losses.

dissimilar to other scenes. To achieve this purpose, we construct a long fixed-size (i,e, P length) window for each shot, in which each shot act as the center shot, i.e., $\mathbf{c}_t = \{\bar{\mathbf{v}}_{t-(P-1)/2}, \cdots, \bar{\mathbf{v}}_t, \cdots, \bar{\mathbf{v}}_{t+(P-1)/2}\}$. And then we simply find pseudo-boundaries by measuring the similarity between shots, i.e, taking the given shot as the center to spread to both sides, and the first shot whose cosine similarity with the center shot is lower the threshold is regarded as a pseudo-boundary, i.e., $\cos(\bar{\mathbf{v}}_j, \bar{\mathbf{v}}_t) \le \mu$. As a result, we divide the fixed-size window of CAT output into three non-overlapping sub-sequences, i.e., $Q_t^{left}, Q_t, Q_t^{right}$. Algorithm details are in the supplementary. The reason for using $\bar{\mathbf{v}}$ to calculate the pseudo-boundary is that adopting the output representations of CAT to calculate the similarity will create more noise, considering the integration of contextual information in the forward process. Related experiments are in the supplementary. Considering the split three sub-sequences as pseudo-scenes, we train the model using InfoNCE loss (van den Oord, Li, and Vinyals 2018):

$$L_{gsm} = -\sum_{\hat{Q}\in\{Q_t^{left},Q_t,Q_t^{right}\}} \log \frac{e^{sim(\hat{\mathbf{v}},\hat{Q})/\tau}}{e^{sim(\hat{\mathbf{v}},\hat{Q})/\tau} + \sum_{Q_r\in\mathcal{N}_r} e^{sim(\hat{\mathbf{v}},Q_r)/\tau}}$$
$$sim(\hat{\mathbf{v}},\hat{Q}) = \cos(\hat{\mathbf{v}}, mean(\hat{Q})) \qquad (5)$$

where $\hat{\mathbf{v}}$ is a randomly sampled shot from $\hat{Q}$, τ is a temperature hyperparameter and $mean(Q)$ means scene-level representations, which utilizes the average pooling of shots in sub-sequence Q. $\mathcal{N}_r$ is the constructed negative scenes using the pseudo-scenes except for $\hat{Q}$, and other pseudo-scenes in the mini-batch.

Local Scene Matching (LSM). Moreover, LSM measures the semantic coherence of the local shots rather than global scene, which learns to decide whether the given two shots belong to the same scene. In detail, we use the center shot $\hat{\mathbf{v}}_t$ as the anchor and construct a tuple $(\hat{\mathbf{v}}_t, \hat{\mathbf{v}}_{pos}, \hat{\mathbf{v}}_{neg})$, where $\hat{\mathbf{v}}_{pos}$ is sampled from Q_t, and $\hat{\mathbf{v}}_{neg}$ is sampled from Q_t^{left} and Q_t^{right}. The loss is defined as:

$$L_{lsm} = -\log \frac{e^{sim(\hat{\mathbf{v}}_t,\hat{\mathbf{v}}_{pos})/\tau}}{e^{sim(\hat{\mathbf{v}}_t,\hat{\mathbf{v}}_{pos})/\tau} + \sum_{\hat{\mathbf{v}}_{neg}\in\mathcal{N}_n} e^{sim(\hat{\mathbf{v}}_t,\hat{\mathbf{v}}_{neg})/\tau}} \qquad (6)$$

where sim denotes cos function. $\mathcal{N}_n$ is the constructed negative shots. In summary, GSM and LSM encourage the CAT to maximize intra-scene similarity, while minimizing inter-scene similarity. The final pre-training loss is defined as:

$$L = L_{smm} + L_{som} + L_{gsm} + L_{lsm}, \qquad (7)$$

Fine-Tuning for Segmentation

As a matter of fact, we have limited videos with boundary labels. Therefore, in the fine-tuning phase, we formulate the video scene segmentation as a binary classification task to identify transitional moments. In detail, given a labeled video $\mathbf{v}$, we develop a scene boundary detection head to infer the boundary prediction for each shot. Following (Mun et al. 2022), we freeze the parameters of the shot encoder f_e, and fine-tune the f_t and boundary detection head. The binary cross-entropy loss can be formulated as:

$$L_p = -\sum_{\hat{\mathbf{v}}_t}[\mathbf{y}_t \log(h_{pre}(\hat{\mathbf{v}}_t)) + (1-\mathbf{y}_t)\log(1-h_{pre}(\hat{\mathbf{v}}_t))] \tag{8}$$

where h_{pre} denotes the boundary detection head. In inferring phase, we predict scene boundary when a shot's prediction score is higher than a pre-defined threshold (i.e., 0.5).

Experiments

Experimental Setups

Dataset. Considering the availability and scale of video segmentation datasets, we adopt the MovieNet dataset following all current state-of-the-art methods (Rao et al. 2020; Chen et al. 2021; Wu et al. 2022b; Mun et al. 2022). MovieNet (Huang et al. 2020) dataset published 1,100 movies where 318 of them are annotated with scene boundaries. In detail, most movies in MovieNet have a time duration between 90 to 120 minutes, providing rich information about individual movie stories. The whole annotation set is split into Train, Validation, and Test sets with the ratio of 10:2:3 on video level following (Huang et al. 2020), the scene boundaries are annotated at shot level. The length of the annotated scenes varies from less than 10s to more than 120s, where the majority last for 10∼30s, more details are in the supplementary.

Comparison Methods. We compare our method CAT with state-of-the-art segmentation approaches: 1) unsupervised methods, i.e., GraphCut (Rasheed and Shah 2005), SCSA (Chasanis, Likas, and Galatsanos 2009), DP (Han and Wu 2011), StoryGraph (Tapaswi, Bäuml, and Stiefelhagen 2014), Grouping (Rotman, Porat, and Ashour 2017). 2) supervised methods, i.e., including Siamese (Baraldi, Grana, and Cucchiara 2015), MS-LSTM (Huang et al. 2020), and LGSS (Rao et al. 2020), and 3) self-supervised methods, including ShotCoL (Chen et al. 2021), SCRL (Wu et al. 2022b), and BaSSL (Mun et al. 2022), more details are in the supplementary.

Evaluation Protocol. Following (Mun et al. 2022), we adopt four commonly used metrics: 1) Average Precision (AP), 2) AUC, 3) F1, and 2) Miou, which measures the averaged intersection over union (IoU) between predicted scene segments and their closest ground truth scene segments.

Shot Feature Encoding. Considering the input features of shots, we construct two modal features following (Rao et al. 2020; Chen et al. 2021; Mun et al. 2022), i.e., the visual and audio modalities, which are encoded independently with separate encoder networks from the input shots. Specifically, visual modality includes place elements to capture the complex semantic information. Place features (2,048 dimensions) are extracted from key-frames in shots with ResNet50 (He et al. 2016). On the other hand, Audio features (512 dimensions) are obtained by concatenating STFT (Umesh, Cohen, and Nelson 1999) features in a shot with a 16K Hz sampling rate and 512 windowed signal length. Multi-modal experiments are in the supplementary.

Implementation Details

For CAT framework, we choose the 2-layer Transformer network with 8 heads, i.e., $N = 8$, as the encoder network architecture. The regression head h_{smm} is a fully connected network with three layers, and the order prediction h_{som} and boundary detection head h_{pre} are fully connected networks with two layers. All weights in the encoder and MLP are randomly initialized. For the pre-training stage, we cross-validate the number of neighbor shots among $L = \{1, 3, 5, 7\}/P = \{13, 15, 17, 19\}$ and $L = 5/P = 17$ is selected due to its good performance and computational efficiency. The optimization method is Adaptive Moment Estimation (Adam), and the learning rate is searched in $\{0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$ to find the best settings for each task. Finally, we set the learning rate as 0.001. The hyper-parameter $\mu = 0.3, \tau = 0.1$. Code is available at https://github.com/njustkmg/CAT.

Video Scene Segmentation

Table 1 summarizes segmentation results against comparison methods. M.S. represents the MovieScenes dataset with 150 annotated movies, M.S-318 denotes the MovieNet dataset with 318 annotated movies, and Eval. means the dataset used for supervised fine-tuning after the self-supervised pre-training. Train., Test., and Val. represent training, testing, and validation sets of MovieNet (Huang et al. 2020). The segmentation results are adopted directly according to the original paper except for * annotation. "-" denotes that the results are not given in original papers. The results reveal that: 1) Supervised methods perform superior to the unsupervised methods, e.g., the LGSS achieves at least 20% improvement in AP. For the reason that supervised data can better guide representation learning. 2) Self-supervised style approaches further improve the performance, which indicates the advantages of pre-training with unsupervised data. 3) CAT achieves the best performance on various criteria, e.g., CAT outperforms the supervised state-of-the-art method, i.e., LGSS by margins of 12.45/4.87 in terms of AP/mIOU, outperforms the state-of-the-art self-supervised method, i.e., BaSSL by margins of 2.15/2.98 /4.91/1.27 in terms of AP/mIOU/F1/AUC, which indicate the effectiveness of context-aware Transformer and pre-training objectives. Besides, CAT using Transformer with only global attention (i.e., CAT with Transformer) performs worse than CAT, revealing the advantages of using context-aware attention. 4) To prevent data leakage, we have reproduced the performance of self-supervised methods on the training dataset (660 movies) for comparison. Compared with other self-supervised approaches that have performance declines, CAT can achieve competitive performance with less training data, with only a decline of 0.33/0.23/0.58/0.18 in terms of AP/mIOU/F1/AUC.

Ablation Study

Impact of individual pretext tasks. To explore the contribution of each pre-training objective, we train models by varying the usage of pre-training tasks. From Table 2, we conclude the following observations: 1) SMM task leads the

W/o SSL	Dataset		AP (↑)	mIOU(↑)	F1(↑)	AUC(↑)
GraphCut	M.S.		14.10	29.70	-	-
SCSA	M.S.		14.70	30.50	-	-
DP	M.S.		15.50	32.00	-	-
Grouping	M.S.		17.60	33.10	-	-
StoryGraph	M.S.		25.10	35.70	-	-
Siamese	M.S.		28.10	36.00	-	-
MS-LSTM	M.S.		46.50	46.20	-	-
LGSS	M.S.		47.10	48.80	-	-
LGSS w/o DP	M.S.		44.90	46.50	38.52	-
LGSS w/o DP*	M.S-318		44.90	46.50	38.52	87.73
W/ SSL	Dataset		AP(↑)	mIOU(↑)	F1(↑)	AUC(↑)
	Pretrain Data	Eval.				
ShotCoL	Train.+Test.+Val.	M.S-318	52.89	-	49.17	-
SCRL	Train.+Test.+Val.	M.S-318	54.82	-	51.43	-
BaSSL	Train.+Test.+Val.	M.S-318	57.40	50.69	47.02	90.54
CAT with transformer	Train.+Test.+Val.	M.S-318	58.35	51.92	50.20	90.59
CAT	Train.+Test.+Val.	M.S-318	**59.55**	**53.67**	**51.93**	**91.81**
ShotCoL	Train. only	M.S-318	48.21	-	46.52	-
SCRL	Train. only	M.S-318	54.55	-	51.39	-
BaSSL*	Train. only	M.S-318	53.36	48.32	43.64	89.26
CAT	Train. only	M.S-318	59.22	53.44	51.35	91.63

Table 1: Scene segmentation results. The compared methods are grouped in two, i.e., (a) approaches that do not use self-supervised learning, including unsupervised and supervised methods, and (b) approaches that adopt self-supervised learning followed by supervised fine-tuning. * denotes our implementations.

Pretext Tasks				Evaluation Metric				
SMM	SOM	GSM	LSM	AP	mIOU	AUC	F1	SUM
✓				16.61	28.00	69.09	21.99	135.69
	✓			39.59	44.34	82.58	38.42	198.93
		✓		44.27	42.03	86.21	32.12	204.63
			✓	29.38	39.88	77.63	30.98	177.87

Table 2: Ablation results for different pre-training tasks.

Local Window	AP	mIOU	AUC	F1
L=3	58.07	53.10	91.15	51.12
L=5	**59.55**	**53.67**	**91.81**	**51.93**
L=7	59.49	53.62	91.77	51.93

Table 3: Ablations of the local encoder window size L. The best scores are in bold.

Contextual Window	AP	mIOU	AUC	F1
P=13	58.73	52.62	91.41	50.44
P=15	59.27	53.39	91.72	51.44
P=17	**59.55**	**53.67**	**91.81**	**51.93**
P=19	59.14	53.46	91.64	51.59

Table 4: Ablations of the contextual window length P in scene-level tasks. The best scores are in bold.

worst performance, which indicates that context-aware pretext tasks (i.e., GSM, LSM, and SOM) can consider contextual relationships well, which is vital for scene segmentation. 2) Scene-level task, i.e., GSM achieves the best performance, which indicates the importance of considering intra-scene and inter-scene distances.

Sensitivity of Hyperparameters. To explore the effect of different hyperparameters: 1) the window size L in the local encoder. 2) the number of local and global attention heads N_l/N_g in context-aware Transformer. 3) the global scene length P in scene-level tasks, we conduct more experiments. More parameter analyses are in the supplementary. Table 3 exhibits the results of parameter L. The performance first increases and then decreases, indicating that a larger window can consider more contextual information, but an oversized one may introduce noisy information. Table 4 records the performance of parameter P, which is similar to parameter L in that a larger global scene can introduce more contextual information, but an oversized one will cause worse

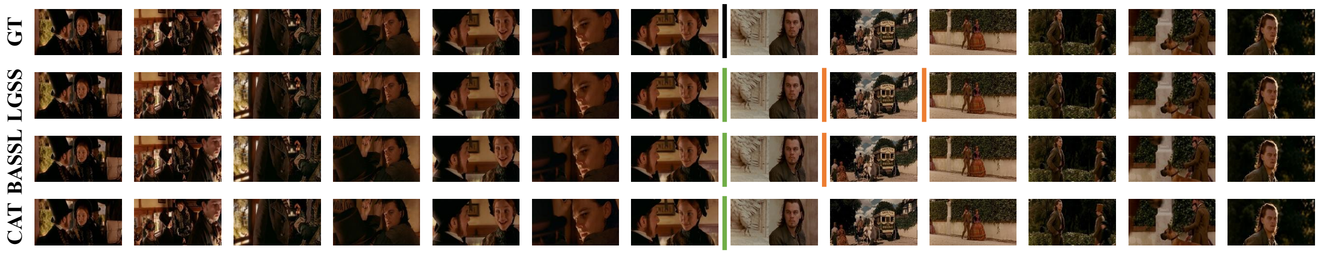

Figure 3: Comparison of boundary detection results from three approaches: LGSS, BaSSL, and our CAT. The green dividing lines indicate the correct boundary, while the yellow lines denote the incorrect ones. "GT" represents the ground truth boundary.

Local-Global Heads		AP	mIOU	AUC	F1
$N_l = 0,$	$N_g = 8$	58.35	51.92	90.59	50.20
$N_l = 2,$	$N_g = 6$	**59.55**	**53.67**	**91.81**	**51.93**
$N_l = 4,$	$N_g = 4$	59.19	53.19	91.66	51.46
$N_l = 6,$	$N_g = 2$	59.10	53.02	91.57	51.08
$N_l = 8,$	$N_g = 0$	58.92	52.27	91.44	50.92

Table 5: Ablations of the local-global number. The best scores are in bold.

performance. Table 5 provides the segmentation results using various local-global heads in CAT. The performance of CAT firstly increases, and then decreases. The CAT acquires the best performance when local-global is (N_l = 2, N_g = 6). The reason may be that the local encoder's window size is limited, thereby the global shots can provide additional supplementary information.

Methods	AP	mIOU	AUC	F1
LGSS (Visual)	39.00	-	-	-
LGSS (Audio)	17.50	-	-	-
LGSS (Visual+Audio)	43.40	-	-	-
ShotCoL (Visual)	46.77	-	-	-
ShotCoL (Audio)	27.92	-	-	-
ShotCoL (Visual+Audio)	44.32	-	-	-
SCRL (Visual)	53.74	-	-	-
SCRL (Audio)	29.39	-	-	-
SCRL (Visual+Audio)	50.80	-	-	-
BaSSL (Visual)	57.40	50.69	90.54	47.02
BaSSL (Audio)	31.69	41.85	79.98	35.49
BaSSL (Visual+Audio)	58.39	52.67	91.09	49.97
CAT (Visual)	59.55	53.67	91.81	51.93
CAT (Audio)	33.41	42.43	80.79	36.40
CAT (Visual+Audio)	**60.20**	**55.49**	**92.17**	**54.78**

Table 6: Comparison results of the multi-modal experiment on MovieNet. Backbones of following methods for each modality are the same.

Performance of Multi-modal Learning

Following (Rao et al. 2020; Yang et al. 2022; Wu et al. 2022b; Mun et al. 2022), we experiment with the proposed method using multi-modal data, i.e., audio and visual modalities. In detail, we adopt the late fusion (i.e., using max pooling of multi-modal predictions) following (Mun et al. 2022). "-" denotes that the results are not given in the original paper. Table 6 records the results, we find that the multi-modal fusion performs better than the single modality, but the audio is a weak modality, which has little promotion.

Visualization

To explore the learning of shot representations, we conduct more experiments. We show segmented cases in Figure 3 to demonstrate the CAT. There are two scenes, we find that a shot with clear change is likely to predict a wrong boundary (i.e., yellow lines) by context limited approaches, even though has similar semantics to contextual shots. However, CAT can successfully predict the boundary. More visualization cases are in supplementary.

Conclusion

In this paper, we study the video segmentation task. With the development of self-supervised learning that adopts both the unsupervised and supervised data into training, we introduce a novel Context-Aware Transformer (CAT) with a self-supervised learning framework to learn high-quality shot representations, for generating well-bounded scene. In detail, CAT utilizes local-global self-attentions to improve the shot encoding. Furthermore, we design shot-to-scene level pretext tasks for learning shot representations, and then we direct fine-tune the CAT with supervised data. Empirical analyses show that CAT can achieve state-of-the-art performance when conducting the scene segmentation task. In the future, how to design a more robust multi-modal fusion strategy is an interesting work.

Acknowledgements

National Key RD Program of China (2022YFF0712100), NSFC (61906092, 62006118, 62276131), Natural Science Foundation of Jiangsu Province of China under Grant (BK20200460), Jiangsu Shuangchuang (Mass Innovation and Entrepreneurship) Talent Program. Young Elite Scientists Sponsorship Program by CAST, CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2021-014B), the Fundamental Research Funds for the Central Universities (NO.NJ2022028, No.30922010317).

References

Ahsan, U.; Sun, C.; and Essa, I. A. 2018. DiscrimNet: Semi-Supervised Action Recognition from Videos using Generative Adversarial Networks. *CoRR*, abs/1801.07230.

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2020. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. In *IJCNN*, 1–8.

Baraldi, L.; Grana, C.; and Cucchiara, R. 2015. A Deep Siamese Network for Scene Detection in Broadcast Videos. In *MM*, 1199–1202. Brisbane, Australia.

Chasanis, V.; Likas, A.; and Galatsanos, N. P. 2009. Scene Detection in Videos Using Shot Clustering and Sequence Alignment. *IEEE Trans. Multim.*, 11(1): 89–100.

Chen, S.; Nie, X.; Fan, D.; Zhang, D.; Bhat, V.; and Hamid, R. 2021. Shot Contrastive Self-Supervised Learning for Scene Boundary Detection. In *CVPR*, 9796–9805. virtual.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, 1597–1607. Virtual.

Cotsaces, C. I.; Nikolaidis, N.; and Pitas, I. 2006. Video shot detection and condensed representation. a review. *IEEE Signal Process. Mag.*, 23(2): 28–37.

Das, A.; and Das, P. P. 2020. Incorporating Domain Knowledge To Improve Topic Segmentation Of Long MOOC Lecture Videos. *CoRR*, abs/2012.07589.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised Visual Representation Learning by Context Prediction. In *ICCV*, 1422–1430. Santiago, Chile.

Han, B.; and Wu, W. 2011. Video scene segmentation using a novel boundary evaluation criterion and dynamic programming. In *ICME*, 1–6. Catalonia, Spain.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778. Las Vegas, NV.

Huang, Q.; Xiong, Y.; Rao, A.; Wang, J.; and Lin, D. 2020. MovieNet: A Holistic Dataset for Movie Understanding. In *ECCV*, volume 12349 of *Lecture Notes in Computer Science*, 709–727. Glasgow, UK.

Jing, L.; and Tian, Y. 2021. Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(11): 4037–4058.

Kuang, H.; Zhu, Y.; Zhang, Z.; Li, X.; Tighe, J.; Schwertfeger, S.; Stachniss, C.; and Li, M. 2021. Video Contrastive Learning with Global Context. *CoRR*, abs/2108.02722.

Li, L.; Chen, Y.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020. HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training. In *EMNLP*, 2046–2065. virtual.

Lin, T.; Liu, X.; Li, X.; Ding, E.; and Wen, S. 2019. BMN: Boundary-Matching Network for Temporal Action Proposal Generation. In *ICCV*, 3888–3897. Seoul, Korea (South).

Lin, T.; Zhao, X.; Su, H.; Wang, C.; and Yang, M. 2018. BSN: Boundary Sensitive Network for Temporal Action Proposal Generation. In *ECCV*, volume 11208, 3–21. Munich, Germany.

Long, F.; Yao, T.; Qiu, Z.; Tian, X.; Luo, J.; and Mei, T. 2020. Learning to Localize Actions from Moments. In *ECCV*, volume 12348, 137–154. Glasgow, UK.

Mun, J.; Shin, M.; Han, G.; Lee, S.; Ha, S.; Lee, J.; and Kim, E. 2022. Boundary-aware Self-supervised Learning for Video Scene Segmentation. *CoRR*, abs/2201.05277.

Pathak, D.; Krähenbühl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context Encoders: Feature Learning by Inpainting. In *CVPR*, 2536–2544. Las Vegas, NV.

Rao, A.; Xu, L.; Xiong, Y.; Xu, G.; Huang, Q.; Zhou, B.; and Lin, D. 2020. A Local-to-Global Approach to Multi-Modal Movie Scene Segmentation. In *CVPR*, 10143–10152. Seattle, WA.

Rasheed, Z.; and Shah, M. 2003. Scene Detection In Hollywood Movies and TV Shows. In *CVPR*, 343–350. Madison, WI.

Rasheed, Z.; and Shah, M. 2005. Detection and representation of scenes in videos. *IEEE Trans. Multim.*, 7(6): 1097–1105.

Roh, B.; Shin, W.; Kim, I.; and Kim, S. 2021. Spatially Consistent Representation Learning. In *CVPR*, 1144–1153. virtual.

Rotman, D.; Porat, D.; and Ashour, G. 2017. Optimal Sequential Grouping for Robust Video Scene Detection Using Multiple Modalities. *Int. J. Semantic Comput.*, 11(2): 193–208.

Rui, Y.; Huang, T. S.; and Mehrotra, S. 1998. Exploring Video Structure Beyond the Shots. In *ICMCS*, 237–240. Austin, Texas.

Sidiropoulos, P.; Mezaris, V.; Kompatsiaris, I.; Meinedo, H.; Bugalho, M.; and Trancoso, I. 2011. Temporal Video Segmentation to Scenes Using High-Level Audiovisual Features. *IEEE Trans. Circuits Syst. Video Technol.*, 21(8): 1163–1177.

Tapaswi, M.; Bäuml, M.; and Stiefelhagen, R. 2014. StoryGraphs: Visualizing Character Interactions as a Timeline. In *CVPR*, 827–834. Columbus, OH.

Umesh, S.; Cohen, L.; and Nelson, D. J. 1999. Fitting the Mel scale. In *ICASSP*, 217–220. Phoenix, Arizona.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR*, abs/1807.03748.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*, 5998–6008. Long Beach, CA.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P. 2008. Extracting and composing robust features with denoising autoencoders. In *ICML*, 1096–1103. Helsinki, Finland.

Wang, Z.; Wu, L.; Li, Z.; Xiong, J.; and Lu, Q. 2021. Overview of Tencent Multi-modal Ads Video Understanding Challenge. *CoRR*, abs/2109.07951.

Wu, H.; Chen, K.; Luo, Y.; Qiao, R.; Ren, B.; Liu, H.; Xie, W.; and Shen, L. 2022a. Scene Consistency Representation Learning for Video Scene Segmentation. *CoRR*, abs/2205.05487.

Wu, H.; Chen, K.; Luo, Y.; Qiao, R.; Ren, B.; Liu, H.; Xie, W.; and Shen, L. 2022b. Scene Consistency Representation Learning for Video Scene Segmentation. *CoRR*, abs/2205.05487.

Xu, D.; Xiao, J.; Zhao, Z.; Shao, J.; Xie, D.; and Zhuang, Y. 2019. Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction. In *CVPR*, 10334–10343. Long Beach, CA.

Yang, Y.; Fu, Z.; Zhan, D.; Liu, Z.; and Jiang, Y. 2021. Semi-Supervised Multi-Modal Multi-Instance Multi-Label Deep Network with Optimal Transport. *IEEE Trans. Knowl. Data Eng.*, 33(2): 696–709.

Yang, Y.; Wu, Y.; Zhan, D.; Liu, Z.; and Jiang, Y. 2018. Complex Object Classification: A Multi-Modal Multi-Instance Multi-Label Deep Network with Optimal Transport. In *KDD*, 2594–2603. London, UK.

Yang, Y.; Zhang, J.; Gao, F.; Gao, X.; and Zhu, H. 2022. DOMFN: A Divergence-Orientated Multi-Modal Fusion Network for Resume Assessment. In *ACMMM*, 1612–1620. Lisboa, Portugal.

Zhu, Y.; Li, X.; Liu, C.; Zolfaghari, M.; Xiong, Y.; Wu, C.; Zhang, Z.; Tighe, J.; Manmatha, R.; and Li, M. 2020. A Comprehensive Study of Deep Video Action Recognition. *CoRR*, abs/2012.06567.