Negative Mining in Self-Supervised Action Localisation

Dezhao Luo
dezhao.luo@qmul.ac.uk
Qilei Li
q.li@qmul.ac.uk
Shaogang Gong
s.gong@qmul.ac.uk

Computer Vision Group,
School of Electronic Engineering and Computer Science,
Queen Mary University of London,
London E1 4NS, UK

Abstract

Self-supervised learning has the potential to explore large-scale unlabelled data for model training. Existing action localisation methods mainly rely on pre-trimmed (pre-segmented) and labelled video clips for model learning. However, in the absence of between-action context, these methods are suboptimal for temporal action localisation (TAL) on *untrimmed* videos. We solve this problem by exploring self-supervised learning for TAL. In particular, we introduce a simple yet effective model, called NEgative MIning in self-supervised action localisation (NEMI), to jointly predict (localisation) content-consistent video fragments, which are considered as activities, and distinguish (classification) them from the other background video content in context. We further explore non-local inter-activity information by training a Transformer-based approach to boundary localisation, which is then adapted to the TAL task. By locating activities and learning to distinguish activities from the context, NEMI can capture the semantic change in a video which is beneficial for TAL in detecting video action boundaries. We evaluate the effectiveness of NEMI by applying the learned model to downstream tasks of temporal action localisation and action detection. Experiments show that NEMI can improve the performance of existing methods by a large margin.

1 Introduction

Temporal action localisation (TAL) [4, 10, 22, 33] has received extensive attention in recent years, given its wide applications in critical scenarios *e.g.* human-computer interaction, intelligent surveillance, and crime tracking. To search certain actions in an untrimmed raw video, previous methods usually train a model to determine the temporal location with boundary labels that are manually annotated. However, the length of videos varies from minutes to hours, plus the number of active instances in a video is unknown *a-priori*, making annotating large-scale untrimmed video action datasets extremely laborious and unscalable. Additionally, because there is no clear definition of the beginning and ending of an action in a video [41], these temporal annotations are prone to be subjective and inconsistent across different annotations, intrinsically offering very noisy labels to model training.

To explore unlabelled raw video data and avoid inconsistent labelling, self-supervised training has attracted increasing attention in video analysis and understanding [6]. Recent attempts have exploited learning video action representation from the order of frames [46], the speed of the video [49], and the similarities in video clip pairs [37]. The general diagram of these methods is to train a model with pre-trimmed videos [15, 18, 35] then adapt it to the downstream action recognition tasks. However, the current widely adopted pre-trimmed datasets contain only action instances in the video while their temporal context is discarded. As a result, these methods are inferior to localising actions in *untrimmed raw videos* that require automatic localisation of action temporal boundaries. In contrast, some other methods [14, 47, 52] have been proposed to dedicatedly localise action boundaries by utilising pseudo labels from pre-trimmed videos, which fails to explore the action context for boundary-sensitive learning. Inspired by [43] which emphasises the importance of negative samples, we argue that the discarded action contexts are equivalently important for boundary learning, as they provide boundaries to be more aligned with the TAL task.

We address these problems by exploring negative mining in self-supervised learning from untrimmed video for the task of TAL. Such untrimmed videos can both minimize human labelling bias and provide richer context information surrounding actions of interest. In particular, we solve two problems: (1) How to generate consistent pseudo-labels for self-supervised learning? (2) How to tailor a pretext task to promote the generalisation ability for TAL? Our solution is formulated as NEgative MIning in self-supervised action localisation (NEMI). Specifically, to address the first problem, we consider a video as a series of activities. Within each activity, we assume its content is relatively consistent. We consider that an activity is a hypothesis of an action in an untrimmed video when labels are unavailable. A model is then trained to predict the location of each activity. We also consider that not only the distinctive characteristics of different actions but also their surrounding context in the video are critical for fine-grained TAL task. Therefore, we further introduce a contrastive loss with negative sample mining to 'pull' closer features of the same activity while to 'push' away features of other activities in context. To address the second problem, we adopt a pre-training and fine-tuning strategy: First, we consider NEMI as a proposal generation model and optimize it with a large-scale dataset without human annotations, as the pre-training. Subsequently, the optimized model is transferred to perform action localisation by fine-tuning on a small-scale dataset. To that end, we explore a Transformer-based approach RTD-Net [36]. RTD-Net extracts video features by an I3D model [3] and adopts a Transformer architecture for direct proposal generation. It has $\approx$ 32M parameters in the Transformer architecture that are not pre-trained. In contrast, the I3D model has $\approx$ 25M [3] pre-trained parameters from large-scale datasets with potentially better generalisation ability.

The contributions of this work are: (1) We make the first attempt to exploit untrimmed videos for TAL model pre-training by exploring rich between-action context information without involving human labels. (2) We devise consistent self-supervised labels by the video activity decomposition and the inter-activity re-assembling strategy. Based on that, the model in trained to complete the activity localisation and negative mining tasks for boundary-sensitive and semantic relationship learning. (3) We consider that an activity is a hypothesis of an action when labels are unavailable, and propose to empower the generalisability of the TAL model by pre-training it with *activity localisation* and transfer it to *action localisation.* We demonstrate the superiority of NEMI over the existing SOTAs on both the temporal action localisation and temporal action detection tasks.

2 Related Works

Temporal Action Localisation. The task of temporal action localisation (TAL) is to localise the temporal boundaries of an action instance. Previous anchor-based methods [9, 10, 13, 51] generate candidates through a length-predefined sliding window. Boundary-based methods [20, 21, 23, 24] firstly predict the boundary probabilities or actionness score for each temporal location and design dedicated matching strategies to form proposal candidates. Subsequent methods, such as CTAP [11], MGG [26], RBRNet [25] and RapNet [8] refine previous methods and propose to adjust the sliding-window proposals with boundary or actionness scores. These methods rely on hand-crafted post-processing designs, and are thus sensitive to noise and unreliable in model generalisation. End-to-end methods [28, 36, 50] aim to predict the boundaries directly. In this regard, Yeung et al. [50] explores reinforcement learning for proposal prediction. AGT [28] proposes to use the encoder-decoder Transformer [39] and build non-linear temporal dependencies for video frames. RTD-Net [36] proposes to replace the encoder with an MLP to address the over-smoothing problem brought about by the feature slowness in videos. RTD-Net [36] trains the transformer-based proposal generator from scratch. In our method, we pre-train the RTD-Net [36] to complete the activity boundary localisation task and transfer it to the TAL task to compare with their training-from-scratch strategy.

Self-Supervised Learning. Self-supervised learning [12] takes advantage of the large-scale unlabelled data for model training. It directly utilises the supervision signal from the data itself, without requiring any tedious labelled data. The trained models are expected to be generalisable and ready to be adapted to a specific downstream task. It has been widely explored for video understanding, with the focus on boundary agnostic and boundary sensitive self-supervised learning.

Boundary Agnostic Self-Supervised Learning with trimmed videos [15, 18, 35] has witnessed great progress in recent years. Early approaches explore video properties for label generation. They usually apply a learnable transformation [27] on the video and learn the feature representations by a visual encoder [34, 38, 45], and then followed by a classifier to predict the transformation parameters. Wang et al. [42] shuffles the frames and takes the original order as the learning target. Fernando et al. [7] designs an odd-one-out network to recognise the unrelated frame in a video sequence. Wei et al. [44] exploits the arrow of time as a supervisory signal. 3D-cubic [19] extends the 2D image jigsaw puzzles [29] to 3D videos. Jing et al. [17] proposes to rotate video clips and use the rotation angle as the learning target. Wang et al. [40] calculates action statistics in a video and train the model by regressing the motion and appearance information. Recently, the inter-video distance has been exploited as a supervisory signal in contrastive learning-based methods [12]. Seco [48] takes the clips generated from the same video as positives and clips from different videos as negatives. IIC [37] further extends the negative samples to intra-negative and inter-negative samples. CVRL [32] explores data augmentations for video representation learning. VideoMoCo [30] improves MoCo [12] for videos by a temporal decay. Even though existing methods are beneficial for the action recognition task, they are boundary agnostic and sub-optimal for the TAL task.

Boundary Sensitive Self-Supervised Learning generates pseudo boundaries from videos as the supervision signal and trains models to perceive these boundaries. Jain et al. [14]

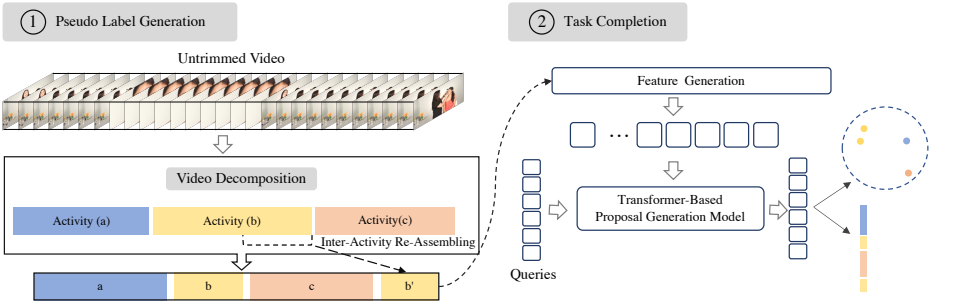

Figure 1: Illustration of the NEMI framework. Firstly, we generate pseudo labels by video decomposition and inter-activity re-assembling. Then we train a model to locate the position of the activities and learn to pull in the same activities and push away the activity and its context.

proposes to generate pseudo labels for action localisation task by abruptly changing detectors to get several atomic actions, and the trained model is then transferred to TAL on untrimmed videos. As a comparison, PAL [52] designs the self-supervision signal by concatenating trimmed videos to detect their boundaries for TAL task-specific pre-training. BSP [47] proposes to synthesise boundaries by concatenating trimmed videos and then train the model to predict the boundary category. Despite their effectiveness, we argue that relying on trimmed videos to form boundaries have the following limitations: (1) The current trimmed video datasets provide weakly-labelled video clip instances with action class label for video clip. These are unreliable/unable to provide fine-grained starting and ending time indices of actions; (2) The inter-action boundaries are semantically different from the action-context boundaries, as the action and its context background usually occur closely. Thus, the model is still insensitive to true boundaries between an action and its context. We propose to solve these problems by introducing contrastive learning to self-supervised learning on untrimmed videos, and providing reliable supervision to learn a generalisable model with extraordinary ability to localise temporal boundaries.

3 Methods

3.1 Problem Definition

Previous works [19, 46, 49] simply design pretext tasks over pre-trimmed video datasets, *e.g.* (UCF101 [35], HMDB51 [15], Kinetics [18]), while overlooking the importance of action boundaries, which is equivalently important for TAL. Recent works [14, 47, 52] use automatically labelled boundaries from pre-trimmed videos to optimise model training. They do not explore any between-action context information for action temporal boundary localisation.

To address the limitations of existing works, we propose to explore natural untrimmed videos for boundary-sensitive learning, which is more generic and practical for real-world applications. Once a discriminative yet generalisable model is trained, we then adapt it to the downstream TAL by task-specific fine-tuning. As shown in Fig.1, we design a multitask learning pipeline to predict the pseudo boundary labels which are generated over *untrimmed* videos. Specifically, we propose pseudo boundaries generation by adaptively decomposing one video into several activities. To make the decomposed activities more discriminative,

we propose an inter-activity re-assembling strategy by injecting a randomly sampled activity instance into the untrimmed video. For task completion, we extract video-level feature representation by utilising 3D-CNN visual encoder and subsequently apply a Transformer-based model to predict the activity boundaries. Given the non-local perception merit of the self-attention modules, the output feature representations can fully consider the inherent relation with other activities and are ready for the multitask predictions, *i.e.* location prediction and contrastive learning with negative mining.

3.2 Pseudo Label Generation

The overall idea of pseudo label generation is to decompose a untrimmed video into several activities and then perform inter-activity re-assembling strategy to obtain the boundaries for model optimisation. The decomposition of untrimmed video is decisively important, as it determines the quality of the pseudo label. In this regard, we propose to use feature-level abrupt change as the criteria for activities localisation. Such a strategy has the advantage of: (1) the content is comparatively consistent within each activity; (2) the activity is temporally scale-invariant, as its length is adapted to the video content [14].

Specifically, suppose that we have an untrimmed video $\mathcal{V}$ from datasets $\mathcal{D}$, where $\mathcal{V} = \{f_t | t \in [1, \mathrm{T}]\}$, and f_t is the t-th frame in $\mathcal{V}$, with T is the total number of frames. We first extract the frame-level representations as $F(f_t)$ and then measure the feature-level variations between each adjacent frames as a criteria to measure the change of the video content. If the measurement of Manhattan distance is larger than the threshold τ, it would be regarded as an activity change between f_i and f_{i+1}, where f_i is the end of the previous activity while f_{i+1} is the beginning of the next activity. The action changes are calculated as:

$$C = \{i | i : |(F(f_{i+1}) - F(f_i)| > \tau\}, \quad i \in [1, T-1] \tag{1}$$

To make the measurement more robust to reflect the activity boundary, we explore two types of features: HOG feature and CNN feature. The HOG feature is dedicated to reflecting image-level inter-frame variation [41] while the CNN feature is in high-level to reflect the semantic variation [14]. Given the pseudo-labelled action boundaries $\mathcal{C}$, the segmented activities in video $\mathcal{V}$ are denoted as $\mathcal{V}_a = \{(C_i, C_{i+1} - 1)\}$, where $i \in \{1, 2, ..., \mathrm{N}\}$ and N is the total number of activities in $\mathcal{V}$. Learning to predict the pseudo label empowers the model to locate the boundaries between activities.

To further encourage the model to distinguish the semantic context, we propose the inter-activity re-assembling strategy by randomly injecting a sampled activity instance into the untrimmed video. Specifically, the activity b' is cropped from one of the existing activities b, as shown in Fig. 1, where b and b' are regarded as positive samples as they display the most similar semantic information, while the other contextual activities are considered as negative samples in contrastive learning.

3.3 Task Completion

We consider NEMI for a boundary localisation task by the two steps: feature extraction and proposal generation.

Feature Extraction The untrimmed video with the pseudo boundary labels is firstly segmented into equally sized temporal intervals called snippets, and then a feature generation

network is applied to extract representation for each snippet. We use an I3D model pre-trained on Kinetics [18] as the feature extractor and freeze the parameters in the following process. A stack of RGB and optical flow frames from each snippet are fed to I3D network to extract the spatial and temporal representation respectively, then they are concatenated to create the final feature.

Proposal Generation To take the non-local information across different activities into consideration, we employ a Transformer-based action localisation model RTD-NET [36] as the baseline. Not only can Transformer blocks build non-linear dependency in the snippet sequence, but also they can predict the action proposals in a direct paradigm. As shown in Fig. 1, we consider NEMI for an activity localisation task, for which the labels are the activity boundaries generated in previous steps (see Section 3.2).

We design two types of localisation strategies with different query initialisations: The first one is called 'activity localisation', where the transformer queries are randomly initialised. The second is 'activity query' as inspired by the image patch query [5] in object detection, where the transformer queries are initialised by activity features.

Considering the objective of the model is to locate an activity, it can be regarded as a binary classification task. We assign the query a binary label by comparing its temporal Intersection over Union (tIoU) towards the target to a given threshold. Then a binary cross-entropy (BCE) loss is introduced to determine if this proposal is an activity, as follows:

$$L_{\text{cls}} = -\frac{1}{K}\sum_{i=1}^{K}(y\log(p_i)+(1-y)\log(1-p_i)), \tag{2}$$

where K is the number of the proposals and p_i is the probability that the i-th proposal is an activity. To predict the boundaries, we also use the localisation loss L_{loc} to minimise the ℓ_1 distance, and the overlap loss L_{overlap} to minimise the tIoU loss as:

$$\begin{aligned} L_{\text{loc}} &= \frac{1}{M}\sum_{i=1}^{M}\left\|\hat{b}_s^i - b_s^i\right\|_{l_1} + \left\|\hat{b}_e^i - b_e^i\right\|_{l_1}, \\ L_{\text{overlap}} &= \frac{1}{M}\sum_{i=1}^{M} 1 - tIoU([\hat{b}_s^i,\hat{b}_e^i],[b_s^i,b_e^i]), \end{aligned} \tag{3}$$

where M is the number of ground truth proposals in the video, $\hat{b}_s^i/\hat{b}_e^i$ denotes the predicted start/end time-stamp of the proposal, and b_s^i/b_s^i is the ground truth.

To provide auxiliary supervision to learn automatic label assignment, we tailor a contrastive loss to maximise the similarity of queries that locate on positive pairs while minimizing those that locate on negative pairs. Given a pair of positive samples in the untrimmed video as a_i and a_j, their corresponding queries are denoted as: $q(a_i)$ and $q(a_j)$. The contrastive loss is formulated as:

$$L_{\text{con}} = -\log\frac{\exp\left(\text{sim}(q(a_i),q(a_j))\right)}{\sum_{k=1,k\neq i}^{M}\exp\left(\text{sim}(q(a_i),q(a_k))\right)} \tag{4}$$

where $\text{sim}(\cdot)$ is the cosine similarity function. The overall learning target is $L = \alpha\cdot L_{\text{cls}} + \beta\cdot L_{\text{loc}} + \gamma\cdot L_{\text{overlap}} + \eta\cdot L_{\text{con}}$.

4 Experiments

In this section, we firstly describe the experimental settings of our methods, then we compare our NEMI with other methods on two tasks: temporal action localisation (TAL) and temporal action detection (TAD). Finally, the ablation study is carried out to validate the effectiveness of model components.

4.1 Experimental Settings

4.1.1 Datasets

We implement our method on two untrimmed action localisation datasets: Thumos14 [16] and ActivityNet-1.3 [2]. These datasets are suitable for our method because they provide natural untrimmed videos.

Thumos14: Thumos14 [16] dataset has 101 classes for action recognition and 20 classes for action detection. It is composed of four parts: training data, validation data, testing data, and background data. In our experiment, we use the validation set for training and the testing set for evaluation which contains 200 and 213 untrimmed videos respectively.

ActivityNet-1.3: ActivityNet-1.3 [2] is a large-scale action detection benchmark, it contains 19,994 videos with 200 action classes annotated. ActivityNet-1.3 is divided into training, validation, and testing sets by a ratio of 2:1:1.

4.1.2 Implementation Details

For video feature extraction, the I3D pre-trained on Kinetics [18] is used as the feature generator. We divide the untrimmed video into equal-length snippets and generate the representation for each. Following the standard protocol [23, 51] in pre-processing, the snippet length is set to 8 for Thumos14, and 16 for ActivityNet-1.3. During the training, the parameters for the visual encoder I3D are frozen. For proposal generation, we choose RTD-Net [36] as the proposal generator. In their implementation of Transformer, they replace the Transformer encoder with a 3-layer MLP. For the Transformer decoder, the number of query patches is set to 32 for Thumos14 and 100 for ActivityNet-1.3, and the number of decoder layers is 6. We use PyTorch [31] to implement the method and use AdamW for optimisation. The batch size is set to 32. The learning rate is 0.0001 and we update the learning rate every 30 epochs by 0.1. The training stops after 100 epochs. For video decomposition, the threshold τ to decide an activity change is set to 0.0715 following [41] for HOG feature and the ratio is set to 0.01 to determine the threshold for CNN feature as [14].

4.2 Comparison with the SOTAs

In this section, we start our experiments by comparing NEMI with other methods. For evaluation, we firstly pre-train a model on the NEMI task and then apply it to two different tasks, temporal action localisation (TAL) and temporal action detection (TAD).

For temporal localisation, we calculate the Average Recall (AR) with Average Number of proposals per video, which are denoted by AR@AN. The experimental results on Thumos14 are summarized in Table 1. With pre-training on NEMI, the AR@50 performance can be improved from 40.1% to 41.7%, AR@100 from 48.3% to 49.7%, which demonstrates our effectiveness. To further evaluate the quality of proposals generated by NEMI, we put the proposals into temporal action detection tasks and use UNet [41] as the proposal classifier. Mean

Method	TAL				TAD				
	@50	@100	@200	@500	0.7	0.6	0.5	0.4	0.3
TURN [9]	21.9	31.9	43.0	57.6	6.3	14.1	24.5	35.3	46.3
CTAP [11]	32.5	42.6	52.0	-	-	-	-	-	-
BSN [23]	37.5	46.1	53.2	60.6	20.0	28.4	36.9	45.0	53.5
MGG [26]	39.9	47.8	54.7	61.4	21.3	29.5	37.4	46.8	53.9
BMN [24]	39.4	47.7	54.7	62.1	20.5	29.7	38.8	47.4	56.0
DBG [20]	37.3	46.7	54.5	62.2	21.7	30.2	39.8	49.4	57.8
RapNet [8]	40.4	48.2	54.9	61.4	-	-	-	-	-
BC-GNN [1]	40.5	49.6	**56.3**	62.8	23.1	31.2	40.4	49.1	57.1
PAL [52]	-	-	-	-	10.9	19.3	30.8	40.3	46.8
RTD-Net [36]*	40.1	48.3	55.6	**62.9**	22.5	34.0	42.8	50.1	56.0
NEMI(ActivityNet-1.3)	41.1	49.2	56.2	62.6	23.3	34.8	44.5	52.2	57.8
NEMI(Thumos14)	**41.7**	**49.7**	**56.3**	62.8	**25.3**	**36.8**	**46.1**	**54.0**	**59.6**

Table 1: Comparison on the TAL task in terms of AR@AN, TAD of mAP@tIoU, on the test set of Thumos14. RTD-Net* denotes our reproduction based on the author released code. NEMI(ActivityNet-1.3) denotes NEMI pre-training on ActivityNet-1.3 and NEMI(Thumos14) on Thumos14.

Method	CTAP [9]	BSN [23]	MGG [26]	BMN [24]	RTD-Net [36]	NEMI
AR@1	-	32.17	-	-	33.05	**33.30**
AR@100	73.17	74.16	74.54	**75.01**	71.70	71.70
AUC	65.73	66.17	66.43	**67.10**	65.78	65.20

Table 2: Comparison between our method with other state-of-the-art proposal generation methods on validation set of ActivityNet-1.3 in terms of AR@AN and AUC.

Average Precision(mAP) with tIoU threshold set [0.3 : 0.1 : 0.7] are calculated and shown in Table 1. When pre-training with NEMI, it obtains 25.3%/36.8%/46.1%/54.0%/59.6%/ on the tIoU of 0.7/0.6/0.5/0.4/0.3 respectively and outperform the RTD-Net [36] by significant margin. These results confirm that proposals generated by NEMI have high quality and work generally well in action detection frameworks.

One can see from Table 1 that pre-training on the large-scale dataset (ActivityNet-1.3) brings less performance gain to that on the small-scale dataset (Thumos14). We argue that this is caused by the domain gap between the datasets used in the pre-training and fine-tuning stages. Even though, pre-training on ActivityNet-1.3 still consistently leads to better performance and outperforms the baseline model, *i.e.* RTD-Net. Table 2 shows that the proposed NEMI can obtain comparable result to RTD-Net, while bringing faster convergence on both large-scale and small-scale datasets, as shown in Fig.3 (a) and (b).

4.3 Ablation Study

In this subsection, we conduct further studies to experimentally investigate the effectiveness of exploring the action context for boundary learning. Both the pre-training and fine-tuning are conducted on the Thumos14 [16] dataset in this subsection. We report the performance on TAD to compare their effectiveness.

Component Analysis As described in Section 3.2, we utilise two features for video decomposition: HOG feature and CNN feature. We start our ablation study by comparing

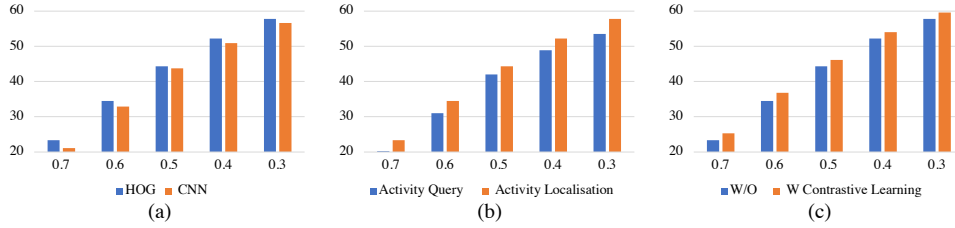

Figure 2: Ablation studies between video decomposition features (a), activity localisation strategy (b) and the contrastive learning (c). The y-axis represents the mAP scores and the x-axis is the tIoU.

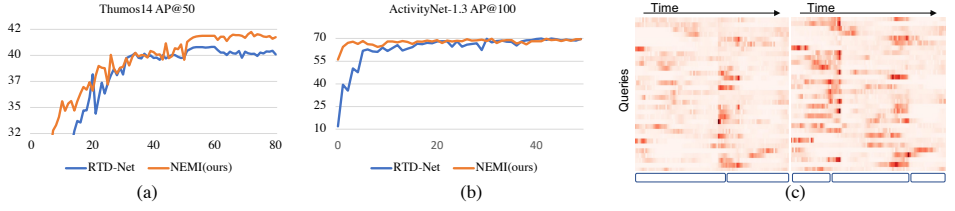

Figure 3: Visualisation of AP learning curve on Thumos14 (a), on ActivityNet-1.3 (b), the transformer attention weight (c).

these two features. The results are summarized in Fig. 2 (a). From the results, we observe although both features give good performance, the HOG feature is better. We attribute this to that the HOG feature considers overall information in each video frame whilst I3D pre-trained by Kinetics focus more on areas of movement only [3]. In the remaining experiments, we choose the HOG feature by default. Then we compare the strategy to locate the activities between activity localisation and activity query in Fig. 2 (b). Also, we add the contrastive learning branch and show its performance in Fig. 2 (c). With contrastive learning, the model yields better performance because it can capture the activity semantics.

Visualisation Fig. 3 (a) shows the AP@50 learning curves on Thumos14. NEMI outperforms RTD-Net with better convergence. The performance is more noticeable after epoch 50. In 3 (b), we show the AP@100 learning curve in the first 50 epochs on ActivityNet-1.3, even though NEMI is comparable to RTD-Net in terms of the performance, it brings faster convergence. To further demonstrate the effectiveness of NEMI, we visualise the normalised multihead attention map of the last Transformer decoder layer in Fig.3. As the warmer the colour denotes the greater the weight, one can see that the queries are more focused on the boundaries.

5 Conclusion

In this work, we introduce a novel negative mining self-supervised learning method referred to as NEMI to explore untrimmed videos for learning temporal action localisation (TAL) task-specific model pre-training. Particularly, we first decompose a untrimmed video into ac-

tivities (hypotheses of actions when labels are unavailable). Then we explore a transformer-based proposal generation model to initialise estimations (proposals) of boundary localisation and optimise it to the TAL task. Experimental results show the effectiveness of NEMI on two downstream tasks: Temporal action localisation and action detection. NEMI is superior because it encourages the model to locate activities by explicitly making contrastive learning against other activities in its surrounding context through negative mining. This is shown to be highly effective in optimising self-supervised learning for temporal action localisation.

References

[1] Yueran Bai, Yingying Wang, Yunhai Tong, Yang Yang, Qiyue Liu, and Junhui Liu. Boundary content graph neural network for temporal action proposal generation. In *ECCV*, pages 121–137. Springer, 2020.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

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

[4] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *CVPR*, pages 1130–1139, 2018.

[5] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *CVPR*, pages 1601–1610, 2021.

[6] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, 2022.

[7] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *CVPR*, pages 3636–3645, 2017.

[8] Jialin Gao, Zhixiang Shi, Guanshuo Wang, Jiani Li, Yufeng Yuan, Shiming Ge, and Xi Zhou. Accurate temporal action proposal generation with relation-aware pyramid network. In *AAAI*, 2020.

[9] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. Turn tap: Temporal unit regression network for temporal action proposals. In *ICCV*, pages 3628–3636, 2017.

[10] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Cascaded boundary regression for temporal action detection. *arXiv preprint arXiv:1705.01180*, 2017.

[11] Jiyang Gao, Kan Chen, and Ram Nevatia. Ctap: Complementary temporal action proposal generation. In *ECCV*, pages 68–83, 2018.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[13] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *CVIU*, 155:1–23, 2017.

[14] Mihir Jain, Amir Ghodrati, and Cees GM Snoek. Actionbytes: Learning from trimmed videos to localize actions. In *CVPR*, pages 1171–1180, 2020.

[15] H Jhuang, H Garrote, E Poggio, T Serre, and T Hmdb. A large video database for human motion recognition. In *ICCV*, volume 4, page 6, 2011.

[16] Yu-Gang Jiang, Jingen Liu, A Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. Thumos challenge: Action recognition with a large number of classes. http://www.thumos.info/, 2014.

[17] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018.

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

[19] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI*, volume 33, pages 8545–8552, 2019.

[20] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Fast learning of temporal action proposal via dense boundary generator. In *AAAI*, 2020.

[21] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *CVPR*, pages 3320–3329, 2021.

[22] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *ACM MM*, pages 988–996, 2017.

[23] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *ECCV*, pages 3–19, 2018.

[24] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*, pages 3889–3898, 2019.

[25] Qinying Liu and Zilei Wang. Progressive boundary refinement network for temporal action detection. In *AAAI*, 2020.

[26] Yuan Liu, Lin Ma, Yifeng Zhang, Wei Liu, and Shih-Fu Chang. Multi-granularity generator for temporal action proposal. In *CVPR*, pages 3604–3613, 2019.

[27] Dezhao Luo, Chang Liu, Yu Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. Video cloze procedure for self-supervised spatio-temporal learning. In *AAAI*, volume 34, 2020.

[28] Megha Nawhal and Greg Mori. Activity graph transformer for temporal action localization. *arXiv preprint arXiv:2101.08540*, 2021.

[29] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016.

[30] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *CVPR*, pages 11205–11214, 2021.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019.

[32] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *CVPR*, pages 6964–6974, 2021.

[33] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *CVPR*, pages 5734–5743, 2017.

[34] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, pages 568–576, 2014.

[35] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[36] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. In *ICCV*, pages 13526–13535, 2021.

[37] Li Tao, Xueting Wang, and Toshihiko Yamasaki. Self-supervised video representation learning using inter-intra contrastive framework. In *ACM MM*, pages 2193–2201, 2020.

[38] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[40] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *CVPR*, pages 4006–4015, 2019.

[41] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, pages 4325–4334, 2017.

[42] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, pages 2794–2802, 2015.

[43] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *AAAI*, volume 36, 2022.

[44] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *CVPR*, pages 8052–8060, 2018.

[45] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018.

[46] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, pages 10334–10343, 2019.

[47] Mengmeng Xu, Juan-Manuel Pérez-Rúa, Victor Escorcia, Brais Martinez, Xiatian Zhu, Li Zhang, Bernard Ghanem, and Tao Xiang. Boundary-sensitive pre-training for temporal localization in videos. In *ICCV*, pages 7220–7230, 2021.

[48] Ting Yao, Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, and Tao Mei. Seco: Exploring sequence supervision for unsupervised representation learning. In *AAAI*, volume 2, page 7, 2021.

[49] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *CVPR*, pages 6548–6557, 2020.

[50] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *CVPR*, pages 2678–2687, 2016.

[51] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *ICCV*, pages 7094–7103, 2019.

[52] Can Zhang, Tianyu Yang, Junwu Weng, Meng Cao, Jue Wang, and Yuexian Zou. Unsupervised pre-training for temporal action localization tasks. *arXiv preprint arXiv:2203.13609*, 2022.