Multiple Instance Triplet Loss for Weakly Supervised Multi-Label Action Localisation of Interacting Persons

Sovan Biswas
University of Bonn, Germany
biswas@iai.uni-bonn.de

Juergen Gall
University of Bonn, Germany
gall@iai.uni-bonn.de

Abstract

With the abundance of videos and the high cost of data annotation, weakly supervised action localisation has gained more attention. However, most of the works on weakly supervised action localisation focus on single action and single person action localisation. Recently, new approaches have been proposed to extend the weakly supervised action localisation task towards multi-label scenarios where multiple persons can interact with each other and perform multiple actions at the same time. For longer videos, these methods subdivide the training videos into very short clips and discard the temporal consistency of actions across these short clips. In this work, we address this issue and propose the Multiple Instance Triplet Loss (MITL) where consistent instances that are temporally close should be more similar than distant and inconsistent instances. It is an extension of the triplet loss to bags, where a bag comprises all person detections at a keyframe. We evaluate our proposed approach on the challenging AVA dataset, where it achieves state-of-the-art results when the weakly labelled training videos are longer than 1 second.

1. Introduction

Humans are multi-tasking by nature, *i.e.*, they perform multiple actions such as reading, sitting, etc. simultaneously. Furthermore, they interact with each other in small groups. For instance, when a person talks the other persons in the group listen. This led to a recent focus on multi-label action recognition datasets like [12, 7, 17, 28] and networks [9, 10] that can be applied to multi-label action detection problems.

One of the main limitations of these existing multi-label action detection approaches lies in the requirement for a high amount of annotated action labels and bounding boxes for action localisation. This requirement of high initial cost and human effort for building suitable models and algorithms limits the usage for large-scale real-world deployment. To circumvent the cost and time associated with data annotation for supervised training, new approaches for weakly supervised multi-label action detection in videos [1, 3] have been proposed. The focus of these approaches is on learning suitable models only from a set of actions that occur in a given video clip as seen in Fig. 1 without any bounding box annotations. These weak annotations are much easier to obtain and reduce the cost and time associated with data annotation drastically. Even though an off-the-shelf person detector can provide very accurate location information, substantial challenges still exist to learn representations for action detection due to the lack of location-action association within the video clip. While this is not an issue for videos that show only one person as in [16], it is very challenging for videos where persons interact in small groups and perform multiple actions at the same time as in [12] and shown in Fig. 1.

Weakly supervised multi-label action detection becomes even more challenging as the length of the video clips increases since the action labels are not provided per keyframe but per video clip. While [3] considered only short video clips of 1 second, [12] proposed a protocol also for longer video clips. In this setting, the chances that an actor leaves or enters the scene, or changes the actions increase over time as shown in Fig. 1. The approach [1] deals with long video clips by subdividing a long video clip into multiple shorter clips. These shorter clips have the same action set annotation as the long video clip during training. This approach, however, discards temporal information like transitions between actions and temporal relations of actions. The accuracy thus decreases rapidly as the clips get longer.

In this paper, we propose a novel approach for weakly supervised multi-label action detection that addresses these limitations for long video clips. The approach is inspired by the triplet loss that aims to learn a representation such that the similarity of a positive instance to an anchor instance is higher than the similarity of a negative instance to the anchor. In the weakly supervised and multi-label setting, however, it is not straightforward to build triplets consist-

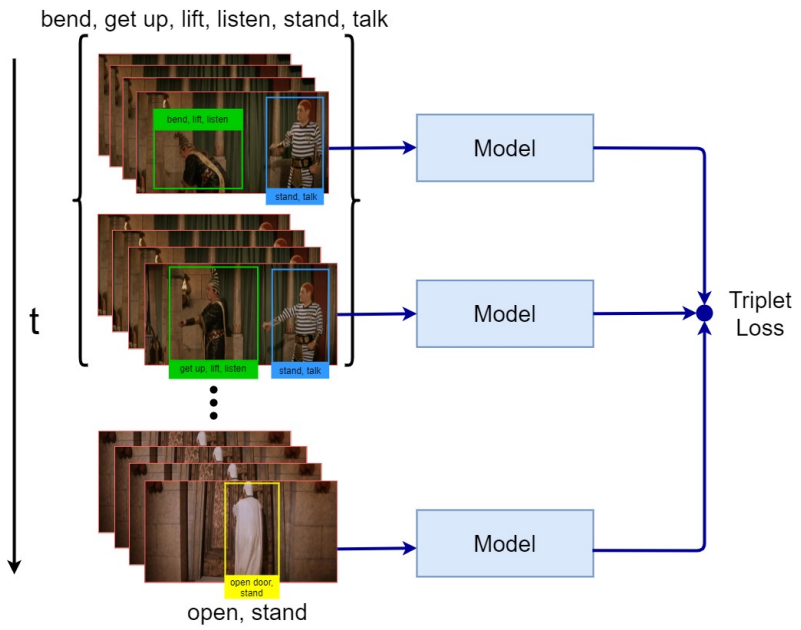

Figure 1: Overview of the Multiple Instance Triplet Loss (MITL) during training. A video shows a scene where two persons are interacting with each other. **Person A** indicated by the green bounding box performs the actions *bend*, *get up*, *lift* and *listen*, where the action *bend* changes to *get up*. **Person B** indicated by the blue bounding box performs the actions *stand* and *talk* consistently over time. These two frames build a positive pair of bags since there is at least one person (**person B**) that performs the same set of actions. For the negative bag, we sample more distant frames where there is at least one person that does not share any action with one person from the first frame. In this example, the **person C** indicated by the yellow bounding box performs the actions *open door* and *stand*, which are not performed by **person B**. Note that the bounding boxes are not provided during training and are only used for visualisation.

ing of an anchor, positive and negative instance as we do not know the corresponding actions of each detected person. We, therefore, extend the triplet loss to bags where a triplet consists of an anchor bag, a positive and a negative bag, and each bag contains all detected instances of a keyframe. The positive bag is from a keyframe that is temporally close to the anchor keyframe whereas the negative bag is from a keyframe that is substantially distant as shown in Fig. 1. For the multiple instance triplet loss, we define the similarities for positive and negative pairs of bags such that the loss selects the instances that are most consistent or least consistent, respectively. For instance in Fig. 1, only person B (blue) of the positive pair of bags performs the same combination of actions in both keyframes while one action of person A (green) changes. For the negative pair, only person C (yellow) and person A do not share any common action while person C and person B both stand. We evaluate the proposed approach on the challenging AVA 2.2 dataset [12], where we outperform the state-of-the-art for weakly supervised multi-label action localisation for training on video clips between 5 and 30 seconds.

2. Related Work

2.1. Spatio-Temporal Action Detection

A widely used strategy for fully supervised spatio-temporal action localisation comprises the joint detection and linking of bounding boxes [11, 29, 32] to form tubelets, which are subsequently used for action classification. Given the considerable improvement of person detectors, many recent methods [9, 36, 8, 17] use them for actor localisation. The focus of these approaches is to learn implicitly or explicitly spatio-temporal interactions between the detected persons. These approaches, however, require for training dense annotations of person locations and action labels, which are often expensive and time-consuming to obtain. This has lead many methods such as [20, 5] to explore weakly supervised learning. Methods such as [30, 19] use multiple instance learning to recognise action characteristics. However, these works assume that a single person performs a single action. In the context of weakly supervised multi-label spatio-temporal action localisation, [1, 3] use multiple instance multi-label learning (MIML) [21, 22, 40] for solving the task. While [1] estimates the uncertainty of

each prediction to mitigate the ambiguity due to multiple labels, [3] iteratively solves the actor-action association to obtain pseudo-labels.

2.2. Contrastive Learning

Contrastive learning, that maximises the intra-class similarity and minimises the inter-class similarity, has been used extensively over the years in various computer vision applications such as image representation learning [4, 14, 24], video representation learning [25, 2], face recognition [27, 6], image captioning [18], phase grounding [13] or future prediction [38]. Even though the objective of contrastive learning largely remained the same, the diverse formulations for intra-class and inter-class similarity has lead to various distinct loss functions such as the triplet loss [15, 27], lifted-structure loss [23], N-pair loss [31], angular loss [34], margin based loss [37], multi-similarity loss [35], circle loss [33] and infoNCE loss [24]. Our approach proposes a loss that is based on the triplet loss [27] but extends it to multiple instance learning in order to deal with the task of weakly supervised multi-label action localisation.

3. Weakly Supervised Multi-Label Action Localisation

The task of multi-label action localisation requires to detect and recognise all actions that are performed by each person in a video as shown in Fig. 1. In contrast to standard action localisation, where it is assumed that one person performs only one action in a video clip, here each person can perform multiple actions at the same time. Further, the actions, as well as the number of persons that are performing these actions, can vary over time as shown in Fig. 1. Nevertheless, we can assume some temporal consistency where at least a subset of the actions is continued in neighbouring frames. Recently, approaches for multi-label action localisation have been proposed that can be trained with weak supervision [1, 3], *i.e.*, without any bounding box annotations for the training videos. In this setting, only the set of actions performed by all persons occurring in a video is provided as annotation. In this work, we follow this protocol for weakly supervised multi-label action localisation, but we focus on learning from temporal consistency.

To exploit the temporal consistency, we propose a triplet loss across time for training. It requires to define triplets (s, p, n) where the positive sample p contains the same actions across time as s and the negative example n contains different actions. Despite the temporal consistency assumption, we cannot assume that all actions are consistent as shown Fig. 1. We also do not know who is performing the given actions in the training video due to the weak supervision. We therefore propose a multiple instance triplet loss that is computed for bags of instances (S, P, N) as shown in Fig. 3, instead of single instances (s, p, n). The loss then aims to minimise the distance between S and P, which will be defined for bags instead of instances, and maximise the distance between S and N. We will first describe the network architecture, which is illustrated in Fig. 2, in Section 3.1 and then describe the novel multiple instance triplet loss, which is illustrated in Fig. 3, in Section 3.2. Finally, we summarise the entire loss function in Section 3.3.

3.1. Network

As the objective is to detect actions, we first generate multiple proposals to locate various actors as in [1, 3]. The proposals are generated using an image-based off-the-shelf person detector based on Faster-RCNN [26] using the ResNeXt backbone [39]. We denote the detected person proposals at time t by $A_t = \{a_i\}$ where a_i is the i^{th} person and A_t is the set of all detected persons at t. As in [12, 10], we then use these person locations to generate person-specific representations from a 3D CNN by applying 2D region of interest pooling at the same spatial location for a temporal window of 1 second as shown in Fig. 2.

These person specific features are passed through two different heads, namely a *classification head* $f(\cdot)$ and a *contrastive head* $g(\cdot)$, as shown in Fig. 2. The classification head consist of a single layer MLP with a sigmoid activation function σ to predict the probabilities for each class, *i.e.*, $f(a_i) = \sigma(W_{class} x_i)$, where x_i is the person specific feature representation from the 3D CNN and $W_{class} \in \mathbb{R}^{D \times N}$ are the weights of the layer. D is the dimension of the feature representation and N is the number of action classes. The contrastive head consists of an MLP with two layers similar to [4], *i.e.*, $g(a_i) = \text{norm}(W_2 \, \text{ReLU}(W_1 x_i))$, where $W_1 \in \mathbb{R}^{D \times 512}$, $W_2 \in \mathbb{R}^{512 \times 512}$, and norm denotes L2 normalisation.

3.2. Multiple Instance Triplet Loss

The objective of contrastive learning is to learn a representation such that similar instances are close to each other, while dissimilar ones are far apart. One common loss for contrastive learning is the triplet loss:

$$\mathcal{L}_{triplet}(s, p, n) = \max(0, \text{sim}(s, n) - \text{sim}(s, p) + \alpha) \quad (1)$$

where $\text{sim}(i, j) = g(a_i)^T g(a_j)$ is the cosine similarity of the L2 normalised embedding for the detected persons a_i and a_j with $-1 \leq \text{sim}(i, j) \leq 1$. α denotes the margin between the positive pair (s, p) and the negative pair (s, n).

In the context of weakly supervised learning, we do not know the labels of the detected bounding boxes a_i and it is therefore not straightforward to generate triplets of instances where a_p contains the same actions as a_s whereas a_n contains different actions. We therefore propose to extend the triplet loss (1) to bags, which contain multiple instances. While we describe in Section 3.2.1 how an anchor bag S and a corresponding positive bag P and negative bag

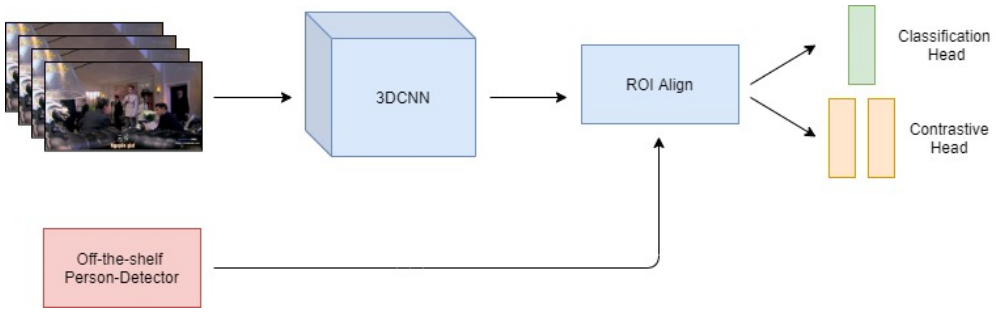

Figure 2: Overview of the network. We use a 3D CNN as backbone and a person detector to get bounding boxes. For each detected bounding box, we extract features from the 3D CNN using region-of-interest pooling. The features are passed through the *classification head* that predicts the class probabilities for the multi-instance and multi-label (MIML) loss and a *contrastive head* that predicts an embedding for the multi-instance triplet (MITL) loss.

N are selected, we define the multiple instance triplet loss for a triplet of bags (S, P, N) by

$$\mathcal{L}_{triplet}(S, P, N) = \max(0, \mathrm{sim_n}(S, N) - \mathrm{sim_p}(S, P) + \alpha) \tag{2}$$

where $\mathrm{sim_n}(S, N)$ defines the similarity for a negative pair of bags and $\mathrm{sim_p}(S, P)$ the similarity for a positive pair. We will describe the similarity measures for bags in Section 3.2.2.

3.2.1 Triplet Selection

In order to build triplets of bags (S_t, P_t, N_t) for each frame t in the training videos, we first take all detections of the frame t as anchors, *i.e.*, $S_t = A_t$. Since persons are likely to perform similar actions over a short period of time as shown in Fig. 1, we select randomly frame $t_p \in \{t-1, t, t+1\}$ and take all detections in frame t_p as positive bag, *i.e.*, $P_t = A_{t_p}$. In case of $t_p = t$, a random transformation like random cropping and mirroring is applied such that P_t is not exactly the same as S_t. For the negative bag N_t, a random frame t_n with $|t_n - t| \geq 100$ is selected and $N_t = A_{t_n}$. Since there is a large temporal gap between N_t and S_t, it is very unlikely that all persons perform the same actions in frame t and t_n.

However, we cannot assume that all persons in S_t and P_t perform the same set of actions over time as shown in Fig. 1. Similarly, it might be possible that persons in S_t and N_t perform some common actions. While it is in principle possible to select negatives from other videos that do not share any actions, this does not provide many negatives for the AVA 2.2 dataset [12] since basic actions like *stand* occur in most frames and videos. We therefore need to define the similarity measures $\mathrm{sim_n}(S, N)$ and $\mathrm{sim_p}(S, P)$ that are robust to missing and overlapping actions, which we will discuss next. For the ease of reading, we omit the frame index t and denote triplets by (S, P, N).

3.2.2 Similarity of Bags

While we cannot assume that each person in S performs the same set of actions in P or is even present in P, we assume that there is at least one close match such that $a_s \in S$ and $a_p \in P$ should be similar as shown in Figs. 1 and 3. For instance, the person in the blue bounding box in Fig. 1 continues the same actions, while the person in the green bounding box changes one of the three actions. In order to measure the similarity between the bags S and P, we therefore only consider the best match. In addition, we also consider that persons performing the same actions are spatially consistent. We thus define $\mathrm{sim_p}$ by

$$\mathrm{sim_p}(S, P) = \max_{a_s \in S, a_p \in P} \left\{ \exp(-\|l_s - l_p\|_2^2) g(a_s)^T g(a_p) \right\} \tag{3}$$

where l_s and l_p are the centroid locations of the detections a_s and a_p, respectively, and $g(a_s)^T g(a_p)$ is the cosine similarity of the L2 normalised embedding for a_s and a_p.

For the negative pair of bags (S, N), we have to consider the possibility that persons $a_s \in S$ and $a_n \in N$ perform the same actions. We therefore assume that there is at least one pair of persons in S and N that do not perform the same combination of actions. We thus define $\mathrm{sim_n}$ by

$$\mathrm{sim_n}(S, N) = \min_{a_s \in S, a_n \in N} g(a_s)^T g(a_n). \tag{4}$$

3.3. Loss Function

In order to train the network, we use besides the multiple instance triplet loss $\mathcal{L}_{triplet}$ two additional loss functions:

$$\mathcal{L} = \mathcal{L}_{miml} + \mathcal{L}_{triplet} + \mathcal{L}_{sim} \tag{5}$$

where $\mathcal{L}_{MIML}$ is a loss for multi-instance and multi-label learning (6) and $\mathcal{L}_{sim}$ is a similarity loss (7) that encourages that the absolute value of two similar samples is high. We add the additional loss functions without weighting them.

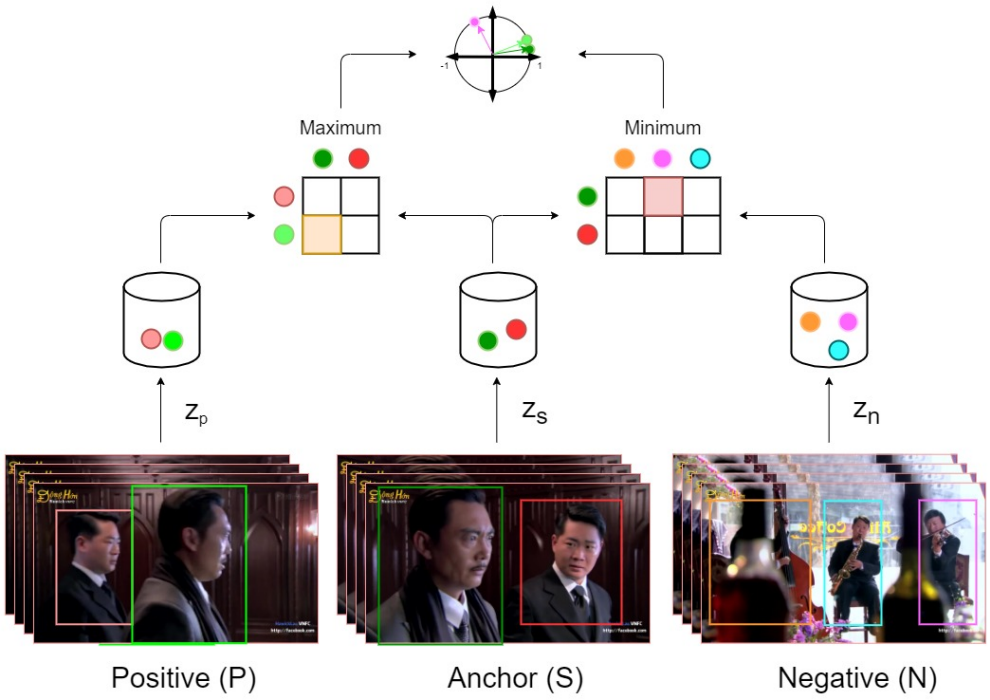

Figure 3: Illustration of the Multiple Instance Triplet Loss (MITL). Any triplet (S, P, N) comprises an anchor S, positive P and negative N bag, where each bag contains a set of detected bounding boxes. As the anchor and positive bag are from the temporal neighbourhood, we assume that they contain at least one instance that is consistent in both bags. This is found by the maximum similarity among the persons in the two bags. The anchor and negative bag are temporally distant. Nevertheless, they can share persons performing the same actions. We thus assume that there is at least one pair of instances that are dissimilar, which is obtained by the minimum similarity. The Multiple Instance Triplet Loss (MITL) thus aims to minimise the distance between the green instances from S and P and to maximise the distance between the green instance from S and the pink instance from N.

The multiple instance triplet loss does not take any supervision into account. Instead we build triplets based on temporal proximity. In order to train the network using the set of labels that is provided for each video clip [1, 3] and contains all actions that are performed by all persons in a video clip, we use a loss for multi-instance and multi-label (MIML) learning [21, 22, 40]. The MIML loss [3] is defined by

$$\mathcal{L}_{miml} = \mathcal{L}\left(Y, \max_i(f(a_i))\right) \tag{6}$$

where Y is a vector that is one for all classes that are present in the video clip and zero otherwise, $f(a_i)$ is the vector that contains the predicted class probabilities for the detection a_i, and $\max$ is the class-wise maximum over all detections. $\mathcal{L}$ is the binary cross entropy.

While the triplet loss encourages the network to learn a representation in order distinguish persons performing similar or dissimilar actions and the margin α can be increased to increase the difference, we show in the experiments that the accuracy decreases when α is too large. For small values of α, however, the absolute similarity for correct pairs can remain low as long as the difference between positive and negative pairs is larger than the margin α. In order to also encourage that the absolute similarity of positive pairs is larger equal to β, we add the similarity loss:

$$\mathcal{L}_{sim} = \max(0, \beta - \mathrm{sim}_\mathrm{p}(S, P)). \tag{7}$$

In the experiments, we evaluate the impact of the loss terms.

4. Experiments

4.1. Dataset

We evaluate our approach on the AVA 2.2 dataset [12]. The dataset contains 235 videos for training and 64 videos for evaluation. Each video is 15 minutes long. The annotation is provided for each person with 60 action classes and bounding box locations at keyframes with a sample rate of

Code: https://github.com/sovan-biswas/MITL

1Hz for the fully supervised setting. The accuracy is measured by mean average precision (mAP) over all actions with an IoU threshold of 0.5 at the key-frames as described in [12]. In the weakly supervised setting, the bounding box information is not used during training. Following the protocol of [1], the training videos are subdivided into $t = 1$, 5, 10, and 30 keyframes, *i.e.*, clips of 1, 5, 10, and 30 seconds, respectively. For each clip, the list of actions that are present in the keyframes is provided. This means that the protocol for $t = 30$ is much more difficult than the protocol for $t = 1$ since we do not know when and where the actions occur in the video clip.

We use Faster RCNN [26] with ResNeXt-101 [39] as backbone to detect persons. This detector is pre-trained on ImageNet and fine-tuned on the COCO dataset. We perform our experiments with SlowFast-50 and SlowFast-101 pre-trained on Kinetics 600. The temporal scope for SlowFast (SF) was set to 64 frames with a stride of 2. We used random cropping and flipping for data augmentation. We randomly crop images of size 224×224 pixels from the resized frame of 256 pixels at its shorter side.

4.2. Comparison to the state-of-the-art

Table 1 shows the comparison of the proposed approach with other state-of-the-art methods. The proposed approach outperforms the state-of-the-art method [1] by $+2.7\%$, $+4.0\%$ and $+5.8\%$ for untrimmed videos of length $t = 5$, 10 and 30 seconds, respectively. This gain can be partly attributed to the better 3D CNN. Thus, we compare the approach with the same Slowfast-50 as [1]. With the same backbone, the proposed approach is still $+1.3\%$, $+1.8\%$ and $+4.2\%$ better for untrimmed videos of length $t = 5$, 10 and 30 seconds, respectively. This gain shows that the proposed approach resolves temporal ambiguity in untrimmed videos. The approach performs better as the length of the untrimmed video increases as seen by the increase in mAP difference compared to [1]. Interestingly, the proposed approach achieves a lower accuracy for trimmed videos of $t = 1$ seconds. With Slowfast-50, it is by -1.7% and -0.9% mAP lower compared to [1] and [3], respectively. For $t = 1$ second, the actions for each keyframe are provided such that a temporal association is not required and the provided labels only need to be assigned to the persons in each keyframe. This is explicitly addressed in [3] and the method thus achieves the highest accuracy for $t = 1$. The approach, however, cannot be directly applied to a setting with $t > 1$. In contrast, the multiple instance triplet loss and the similarity loss are only suitable for a setting with $t > 1$.

4.3. Ablation Studies

4.3.1 Impact of Loss Functions

In (5), we use the loss functions $\mathcal{L}_{miml}$, $\mathcal{L}_{triplet}$ and $\mathcal{L}_{sim}$. While $\mathcal{L}_{miml}$ is always required since it is the only loss function that takes the weak annotations into account, Table 2 shows the quantitative impact of the other loss functions for $t = 5$ and 30 seconds, respectively. $\mathcal{L}_{triplet}$ increases mAP by $+0.5\%$ at $t = 30$ seconds, indicating the impact of contrastive learning in removing temporal ambiguity in long untrimmed videos. When both $\mathcal{L}_{triplet}$ and $\mathcal{L}_{sim}$ are used, mAP increases by $+0.7\%$ and $+1.2\%$ mAP for $t = 5$ and 30 seconds, respectively. $\mathcal{L}_{sim}$ ensures that the absolute similarity of positive matches over time is large.

4.3.2 Impact of α

In (2), α causes the model to separate negative and positive bags, and as α increases the distance becomes larger. Due to the multi-label scenario, persons often share some actions as shown in Fig. 1. If a positive pair does not share all actions or a negative pair shares some actions, a large value of α can have a negative impact. In Table 3, we thus quantitatively analyse the impact of α using SlowFast-50 as backbone for the $t = 30$ setting. As expected, the accuracy decreases for $\alpha = 0.3$ and the best mAP of 15.6% is obtained for α between 0.05 and 0.1. In all other experiments, we use $\alpha = 0.05$. Interestingly, the approach is able to classify the least 10 frequently occurring classes better with $\alpha = 0.3$ as seen by an increase of mAP to 2.8%. This is in contrast to the deteriorating performance of the top 5 frequently occurring classes with an increasing value of α. This is due to the fact that frequently occurring classes are more likely to be present both in an anchor and negative bag and are thus more likely to be shared by a negative pair.

4.3.3 Impact of β

The loss (7) encourages that the absolute similarity of positive pairs is larger equal to β. While we evaluated the impact of the loss already in Table 2, we evaluate the impact of β in Table 4 for $t = 30$ seconds. In case of $\beta = 1$, the loss aims to maximise the absolute similarity since $\mathrm{sim}_{\mathrm{p}}$ cannot be larger than 1. This reduces the accuracy compared $\beta = 0.8$, which we use in our experiments.

4.3.4 Variants of $\mathrm{sim}_{\mathrm{n}}$

As we discussed in Section 3.2.2, we assume that there is at least one pair of persons in the anchor and negative bag that do not perform the same combination of actions and thus take the minimum over all pairs to compute $\mathrm{sim}_{\mathrm{n}}$ (4). If we take the maximum instead of the minimum, we would

Table 1: Comparison of the proposed method with other state-of-the-art methods. The clip length of the training videos with weak annotations is denoted by $t = 1$, 5, 10, and 30 seconds. The larger t is, the more difficult is the task.

Methods	3D-CNN	$t = 1$	$t = 5$	$t = 10$	$t = 30$
Uncertainty-Aware[1]	SF-50	22.4	18.0	15.8	11.4
Actor-Action[3]	SF-50	21.6	-	-	-
Actor-Action[3]	SF-101	**25.1**	-	-	-
Proposed Approach	SF-50	20.7	19.3	17.6	15.6
Proposed Approach	SF-101	22.1	**20.7**	**19.8**	**17.2**

Table 2: Impact of loss functions for $t = 5$ and 30 seconds. $\mathcal{L}_{miml}$ is the MIML loss, $\mathcal{L}_{triplet}$ is the multiple instance triplet loss and $\mathcal{L}_{sim}$ is the similarity loss. **Y** denotes that the loss function has been used.

$\mathcal{L}_{miml}$	$\mathcal{L}_{triplet}$	$\mathcal{L}_{sim}$	$t = 5$	$t = 30$
Y	-	-	18.6	14.4
Y	**Y**	-	-	14.9
Y	**Y**	**Y**	**19.3**	**15.6**

Table 3: Impact of the margin α for $t = 30$ seconds. The Least 10 and Top 5 indicate the least 10 frequently and top 5 frequently occurring classes in the training dataset.

α	0	0.05	0.1	0.3
Overall	15.1	**15.6**	15.6	14.5
Least 10	1.7	1.9	2.7	**2.8**
Top 5	55.4	**55.4**	55.3	54.2

Table 4: Impact of the margin β for $t = 30$ seconds.

β	1	0.8
Overall	13.9	**15.6**

assume that none of the pairs shares any action. In Table 5, we compare the maximum and the minimum using SlowFast-50 as backbone for the $t = 30$ setting. The approach achieves mAP of 15.6% for the min and 14.9% for the max operation. Taking the minimum for the negative pairs of bags also outperforms the maximum for the least 10 and top 5 occurring classes. However, the gain of min is larger for the top 5 occurring classes since they are more likely to be shared by a negative pair.

5. Conclusion

In this paper, we proposed a novel approach based on contrastive learning for weakly supervised multi-label action localisation. In this setting, only the list of actions occurring in each training video is provided. So, in order to learn a better representation despite weak annotation, we introduced the novel Multiple Instance Triplet Loss (MITL) which takes the similarity of bags instead of instances into account. Later, we evaluated our proposed approach on the challenging AVA dataset, where it is difficult to define negative pairs since some actions like standing occur very frequently. We therefore addressed this issue by defining different similarity functions for positive and negatives bags. For the setting where the training videos are longer than 1 second, our approach achieved state-of-the-art results.

Table 5: Comparison of taking the min or max to compute sim_n (4) for $t = 30$ seconds. Least 10 and Top 5 indicate the least 10 frequently and top 5 frequently occurring classes in the training dataset.

Negative-Bag Similarity	max	min
Overall	14.9	**15.6**
Least 10	1.7	**1.9**
Top 5	54.7	**55.4**

Acknowledgement

The work has been financially supported by the ERC Starting Grant ARCA (677650).

References

[1] Anurag Arnab, Chen Sun, Arsha Nagrani, and Cordelia Schmid. Uncertainty-Aware Weakly Supervised Action Detection from Untrimmed Videos. In *ECCV*, pages 751–768, 2020. 1, 2, 3, 5, 6, 7

[2] Nadine Behrmann, Jurgen Gall, and Mehdi Noroozi. Unsupervised video representation learning by bidirectional feature prediction. In *WACV*, pages 1670–1679, 2021. 3

[3] Sovan Biswas and Juergen Gall. Discovering Multi-Label Actor-Action Association in a Weakly Supervised Setting. In *ACCV*, 2020. 1, 2, 3, 5, 6, 7

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 3

[5] Guilhem Chéron, Jean-Baptiste Alayrac, Ivan Laptev, and Cordelia Schmid. A flexible model for training action localization with varying levels of supervision. In *NeurIPS*, pages 942–953, 2018. 2

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 3

[7] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *ECCV*, pages 593–610, 2020. 1

[8] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 2

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast Networks for Video Recognition. In *ICCV*, pages 6202–6211, 2019. 1, 2

[10] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video Action Transformer Network. In *CVPR*, pages 244–253, 2019. 1, 3

[11] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *CVPR*, pages 759–768, 2015. 2

[12] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. AVA: A Video dataset of Spatio-Temporally Localized Atomic Visual Actions. In *CVPR*, pages 6047–6056, 2018. 1, 2, 3, 4, 5, 6

[13] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *ECCV*, pages 752–768, 2020. 3

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 3

[15] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pages 84–92. Springer, 2015. 3

[16] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, 2013. 1

[17] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, pages 10236–10247, 2020. 1, 2

[18] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data. In *ECCV*, pages 338–354, 2018. 3

[19] Pascal Mettes and Cees GM Snoek. Spatio-temporal instance learning: Action tubes from class supervision. *arXiv preprint arXiv:1807.02800*, 2018. 2

[20] Pascal Mettes, Cees GM Snoek, and Shih-Fu Chang. Localizing actions from video labels and pseudo-annotations. In *BMVC*, 2017. 2

[21] Cam-Tu Nguyen, De-Chuan Zhan, and Zhi-Hua Zhou. Multi-modal image annotation with multi-instance multi-label lda. In *IJCAI*, 2013. 2, 5

[22] Nam Nguyen. A new SVM approach to multi-instance multi-label learning. In *ICDM*, pages 384–392, 2010. 2, 5

[23] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, pages 4004–4012, 2016. 3

[24] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[25] 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. 3

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*, pages 91–99, 2015. 3, 6

[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 3

[28] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *ECCV*, pages 510–526, 2016. 1

[29] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip Torr, and Fabio Cuzzolin. Online Real-Time multiple spatiotemporal action localisation and prediction. In *ICCV*, pages 3657–3666, 2017. 2

[30] Parthipan Siva and Tao Xiang. Weakly Supervised Action Detection. In *BMVC*, page 6, 2011. 2

[31] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, pages 1857–1865, 2016. 3

[32] Lin Song, Shiwei Zhang, Gang Yu, and Hongbin Sun. TACNet: Transition-Aware Context Network for Spatio-Temporal Action Detection. In *CVPR*, pages 11987–11995, 2019. 2

[33] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pages 6398–6407, 2020. 3

[34] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *ICCV*, pages 2593–2601, 2017. 3

[35] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR*, pages 5022–5030, 2019. 3

[36] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, pages 284–293, 2019. 2

[37] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *ICCV*, pages 2840–2848, 2017. 3

[38] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. Learning to anticipate egocentric actions by imagination. *IEEE Transactions on Image Processing*, 30:1143–1152, 2020. 3

[39] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 3, 6

[40] Hao Yang, Joey Tianyi Zhou, Jianfei Cai, and Yew Soon Ong. MIML-FCN+: Multi-instance multi-label learning via fully convolutional networks with privileged information. In *CVPR*, pages 1577–1585, 2017. 2, 5