Query by Activity Video in the Wild

Tao HU   William Thong   Pascal Mettes   Cees G.M. Snoek
University of Amsterdam

Abstract

This paper focuses on activity retrieval from a video query in an imbalanced scenario. In current query-by-activity-video literature, a common assumption is that all activities have sufficient labelled examples when learning an embedding. This assumption does however practically not hold, as only a portion of activities have many examples, while other activities are only described by few examples. In this paper, we propose a visual-semantic embedding network that explicitly deals with the imbalanced scenario for activity retrieval. Our network contains two novel modules. The visual alignment module performs a global alignment between the input video and fixed-sized visual bank representations for all activities. The semantic module performs an alignment between the input video and fixed-sized semantic activity representations. By matching videos with both visual and semantic activity representations that are of equal size over all activities, we no longer ignore infrequent activities during retrieval. Experiments on a new imbalanced activity retrieval benchmark show the effectiveness of our approach for all types of activities.

1. Introduction

This paper investigates the problem of activity retrieval given a video as example. In current literature, activity retrieval is more often framed as a classification task [19], a localization task [13], or as a retrieval-by-text problem [8]. For the less common task of activity retrieval by video, several works have shown that activities can be retrieved directly from a user-provided query [5, 7, 33, 38, 39]. A standard assumption however is that the activities form a closed set, *i.e.* they assume a fixed set of activities, each with many training videos. In practice, most activities will have few examples. Without an explicit focus on such activities, they will be ignored in favour of activities with many examples. In this work, we focus on learning balanced video representations of activities for retrieval, regardless of whether they have many or few examples.

Learning with imbalanced data is an active research topics for various visual tasks, including image classification [22, 47], image segmentation [2, 20], and object de-
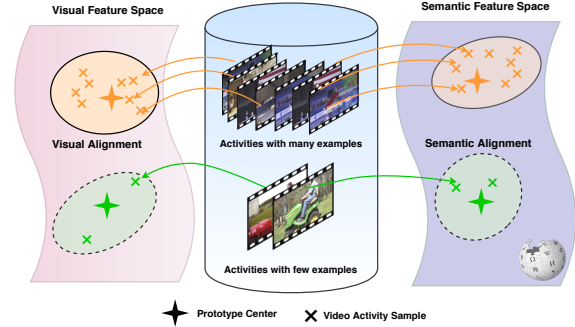

Figure 1: **Our motivation.** We aim at retrieving activities by an activity video query. The training set is composed of activities with many examples and activities with few examples. We propose a visual-semantic alignment to balance the retrieval performance between base and novel classes.

tection [28]. A central theme in these works is to either make classes with few examples more prominent, or switch to a setting where all classes have an equally-sized representation, *e.g.* using memory banks [23, 46] or prototypes [24, 37]. Here, we take inspiration from existing imbalanced tasks for the problem of imbalanced activity retrieval from video queries. We seek to introduce two alignment modules to match the activity feature regardless of whether they have many or few examples, see Figure 1. Different from current works, we do so by using both visual and semantic prototypes, where we emphasize the importance of a global alignment with respect to all activities. This allows us to better focus on equally-sized activity representations during training, which in turn results in a more balanced retrieval.

As a first contribution in this work, we introduce a new task about video query by activity in the wild and emphasize the importance of performance balance between the activities with many examples and activities with few examples. Second, we introduce a visual-semantic embedding network for retrieval by a video query. The network extends the standard classification loss in deep networks with two novel modules. The visual alignment module maintains a visual bank with equal space for each activity. The representation of an input video is globally aligned with the visual bank representations of all activities to obtain a loss

that disregards the amount of examples available for each activity. The semantic alignment module performs a similar global alignment and loss, but between the input video and video-independent semantic representations of activities such as word embeddings. These modules explicitly target the problem of imbalance in video dataset. Third, we reorganized the ActivityNet dataset [3] to emulate an imbalanced retrieval dataset, along with new data splits and example sampling. we perform extensive evaluation and analyses to examine the workings of our approach for imbalanced activity retrieval. Lastly, we show our ability on video clips [25] and moments [8].

2. Related work

2.1. Video retrieval

For video retrieval, one common direction is to retrieve videos by a textual query [13, 10, 14, 25, 27, 29, 40, 45]. Hendricks *et al.* [13] propose a network that localizes text queries in videos using local and global temporal video representations. Hendricks *et al.* [14] further propose to model the context as a latent variable to bridge the gap between videos and textual queries. Beyond video retrieval, a number of recent works have investigated localized retrieval from text queries. Notably, Gao *et al.* [10] and Miech *et al.* [25] jointly model text and video clips in a shared space to obtain fixed-length local videos clips as retrieval output. Similar endeavours have been proposed to retrieve localized video moments from untrimmed datasets given a text query [27]. In this work, we also extend the retrieval beyond videos only to clips and moments, but do so by using an input video as query, rather than text.

For activity retrieval by query video, current works are generally concerned with an efficient matching setup between query and test videos. Examples include retrieval using hashing [39] and retrieval using quantized video representations [5]. A common starting assumptions is that the activities to retrieve have ample training examples to learn such an efficient matching. In this work, we challenge this assumption and propose a network is retrieve both activities with many examples and activities with few examples from a query video.

2.2. Learning with imbalanced data

When dealing with frequent classes (base classes [11]) and infrequent classes (novel classes [11]), a persistent issue is overfitting to the base classes. Transfer learning to novel classes provides a way to boost the performance on novel classes, although this is paired a catastrophic forgetting problem on base classes [35]. Meta-learning is similarly focused on improving generalization to novel classes, *e.g.* through a few steps of fine-tuning [9, 34]. However, these methods only consider the generalization on novel classes while ignoring the performance of base classes [23]. We aim to achieve a balance between both.

To attain such a balance, early work attempted to use a single example of the novel class to adapt classifiers from similar base classes using hand-crafted features [1]. Learning with imbalanced classes has since actively been researched in image classification [22, 47], image segmentation [2, 20], and object detection [28]. Bharath *et al.* [11] for example tackle the imbalance problem by hallucinating additional training examples for rare novel classes. As an extension of these works, we explore the balancing of base and novel classes for the problem of activity retrieval by a video example.

3. Visual-Semantic Embedding Network

We aim to learn video representations for activity retrieval, where the task is to retrieve videos of the same activity given a query video. Let $\{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ be a set of N activity videos, where x is a video of T frames describing an activity $y \in \mathcal{Y}$. Our goal is to learn an embedding function $f(\cdot) \in \mathbb{R}^C$ such that two different videos $x^{(i)}$ and $x^{(j)}$ of the same activity y are close in the embedding space.

In large collections of activities, there usually exists an imbalance in the number of examples per activity. Following Hariharan and Girshick [11], we denote activities with many examples as the *base* classes and activities with few examples as the *novel* classes. Formally, $\mathcal{Y}$ is then split into $\mathcal{Y}_{base}$ and $\mathcal{Y}_{novel}$, with $\mathcal{Y}_{base} \cap \mathcal{Y}_{novel} = \emptyset$. Having an imbalanced training set causes the embedding function $f(\cdot)$ to be geared towards $\mathcal{Y}_{base}$ in the evaluation phase. As a consequence, this induces a poor retrieval performance for the under-represented classes $\mathcal{Y}_{novel}$. To alleviate this issue, we propose two alignment modules to preserve the *visual* and *semantic* representations of all activities.

First, we describe how to learn activity representations for all classes with a simple classification network (Section 3.1). Second, we introduce two alignment modules to better handle novel classes. We propose a visual alignment module to preserve the activity representations over time (Section 3.2), and a semantic alignment module to enforce activity representations to be semantically meaningful (Section 3.3). Finally, we show how to train and evaluate the overall model (Section 3.4). Figure 2 illustrates the proposed Visual-Semantic Embedding Network.

3.1. Action representations

To learn video representations of activities, we opt for a frame-level convolutional network (ConvNet) as an embedding function. Working at the frame-level rather than at the video level (*e.g.* with 3D convolutions) offers more flexibility at the evaluation phase. In this work, frame-level representations enable us to perform localized retrieval, *e.g.* retrieval of video clips [25] or video moments [8].

We extract the embedding representation for every frame x_t and simply average them over time to obtain a video-
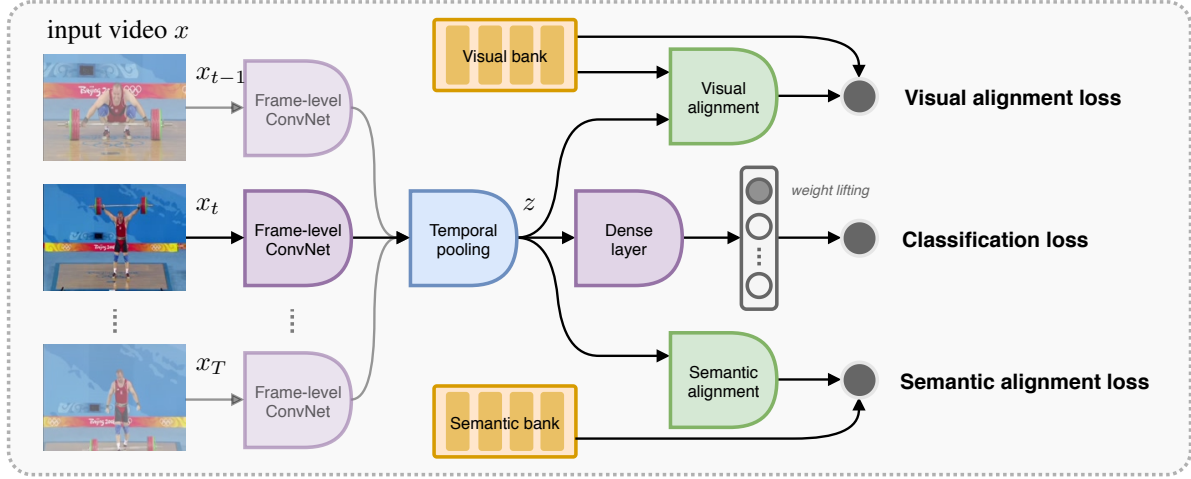
Figure 2: **Visual-Semantic Embedding Network**. For training phase, we adopt the visual alignment loss, semantic alignment loss, classification loss. For testing phase, we only use the frame-level embedding z. The arrow indicates the direction of information flow.

level representation $z \in \mathbb{R}^C$:

$$z = \frac{1}{T} \sum_{t=1}^{T} f(x_t). \qquad (1)$$

The embedding representation is then further projected on a label space for classification. The probability of class $c \in \mathcal{Y}$ given an embedding representation z is:

$$p_A(y = c | z) = \frac{\exp(-W_c \cdot z)}{\sum_{k \in \mathcal{Y}} \exp(-W_k \cdot z)}, \qquad (2)$$

where W is the learnable parameter of the linear projection.

3.2. Visual alignment

While a standard classification embedding uses examples of all activities, the loss is in practice dominated by activities with many examples. In an effort to balance base and novel activity representations, we first focus on a visual alignment between all activities. Let $V \in \mathbb{R}^{K \times C}$ denote a visual bank matrix consisting of features representations of dimension C for every activity $y \in \mathcal{Y}$. The size of the bank then corresponds to $K = |\mathcal{Y}|$ activities. The idea of the visual bank is to obtain a single prototypical representation for every activity. Hence, all activities are treated equally, regardless of the number of examples available for training. For a new activity embedding z of activity y, we update the visual bank V through a convex combination of the current embedding representation z and the corresponding entry in V followed by an ℓ_2 normalization:

$$\begin{aligned} V_y &= \alpha \frac{z}{\|z\|_2} + (1 - \alpha) V_y, \\ V_y &= V_y / \|V_y\|, \end{aligned} \qquad (3)$$

where α controls the amount of update in the visual bank. The visual bank is initialized to zero when training.

Building upon such visual banks, we propose to align the representations of different activities. For this purpose, we rely on the attention mechanism from the non-local operator [41]. Compared to the original non-local block, we aim to capture the relation between different prototypical representations rather than spatial [41, 16] or temporal [41, 42] relations. The structure of the visual alignment module is illustrated in Figure 3.

Let GA denote a global alignment operator between the visual bank representation of one activity in V, *i.e.* and $\text{GA} : (\mathbb{R}^{1 \times C}, \mathbb{R}^{K \times C}) \mapsto \mathbb{R}^{1 \times C}$. When compared to the prototypes in the visual bank, the aligned representation $z^\star = \text{GA}(V_c, z)$ can then be used to provide the probability of class c:

$$p_V(y = c | z) = \frac{\exp\left(-d(z^\star, V_c)/\tau\right)}{\sum_{k \in \mathcal{Y}} \exp\left(-d(z^\star, V_k)/\tau\right)}, \qquad (4)$$

where τ is the temperature of the softmax function and d is the Euclidean distance.

3.3. Semantic alignment

We additionally leverage word embeddings of activity names as a prior information. A semantic representation of an activity encapsulates relations amongst all pairs of activity classes. We use this information to additionally align activity representations towards such semantic knowledge. We denote $\phi(y) \in \mathbb{R}^W$ as the word embedding of the activity y. Let $S \in \mathbb{R}^{K \times W}$ be the semantic bank which compiles the word embedding $\phi(y)$ of all K activities. For the semantic alignment, we simply opt for a multilayer perceptron $g(\cdot)$. Similar to the visual alignment, a probability for class c can be derived after aligning the representation z
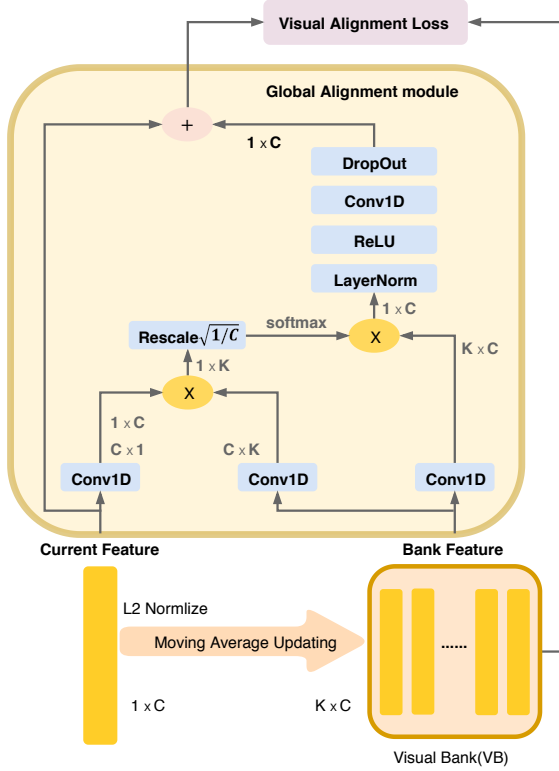
Figure 3: **Visual Alignment Module**. K means class number, C denotes feature dimension. For every video feature with label y, we update the Visual Bank in the according y index. Then current feature and bank feature are fed into Global Alignment part to balance the feature from all classes. Finally a visual alignment loss is applied. Best viewed in color.

with the semantic activity embedding:

$$p_{\mathrm{S}}(y = c|z) = \frac{\exp\big(-d(g(z), S_c)/\tau\big)}{\sum_{k \in \mathcal{Y}} \exp\big(-d(g(z), S_c)/\tau\big)}, \quad (5)$$

Compared to the visual bank, the semantic bank remains fixed during training. We initialize the semantic bank from an existing word embedding (*e.g.* word2vec [26]).

3.4. Optimization

Training the overall network amounts to minimizing the cross-entropy loss function for all three components over the training set:

$$\mathcal{L} = -\log(p_A) - \lambda_V \log(p_V) - \lambda_S \log(p_S), \quad (6)$$

where λ_V and λ_S are trade-off hyper-parameters. Once the network has been trained, we extract the video-level representations z followed by an ℓ_2 normalization for all videos in the gallery set. Similarities among videos are then measured with the Euclidean distance.

Table 1: **VR-ActivityNet statistics** for video retrieval. C_{0-100} are base classes with many examples, $C_{100-200}$ are novel classes with few examples. Some distractors with irrelevant content are constructed to simulate the real-life retrieval scenario in the testing phase.

		train	validation	test
Base	C_{0-100}	6095	1000	3597
Novel	$C_{100-120}$	100	200	-
	$C_{120-200}$	400	-	2800
#distractor		-	-	573
Total		6595	1200	6970

4. Experimental Setup

Implementation details We employ ResNet-18 [12] as a backbone network with weights pre-trained on ImageNet [6]. Fine-tuning is done by the Adam [21] optimizer on one Nvidia GTX 1080TI. We set the learning rate to $1e-4$ with a weight decay of $1e-5$ for 16k iterations and reduce the learning rate to $1e-5$ after 8k iterations. We use a batch size of 16. The trade-off hyper-parameters λ_V, λ_S are set to 1 by cross-validation and the convex coefficient α of the visual bank update is set to 0.9. Three video frames are extracted per second, resulting in an average of 32 frames per activity video. Every frame is randomly cropped and resized to 112×112. We use ELMo [32] with 1,024 dimension as our default word embedding method. We use PyTorch [30] for implementation and the Faiss [17] library to measure video similarities.

VR-ActivityNet We reorganize ActivityNet1.3 [3] for our video retrieval and called the reorganization VR-ActivityNet. As our method aims at evaluating the performance of *base* classes and *novel* classes, we split the 200 activity labels into 100 base classes (C_{0-100}) and 100 novel classes ($C_{100-200}$). We also divide the dataset into training, validation, testing set. The validation set is trying to evaluate the balance performance between the C_{0-100} and $C_{100-120}$. Similarly, the testing set is designed to evaluate the balance performance between the C_{0-100} and $C_{120-200}$. Detailed activity splits are shown in the supplementary file.

We split 10,024 untrimmed long videos from ActivityNet training set into trimmed meaningful activity segments and randomly generate a number of meaningless distractor segments. We then formulate the training and validation set of VR-ActivityNet. We utilize 4,926 untrimmed long videos from the ActivityNet validation-set to generate the trimmed segments of our testing set in VR-ActivityNet. The number of activity videos per subset in VR-ActivityNet is shown in Table 1. For novel classes in the training data, only 5 samples per novel class are accessible. For validation and testing data, the sample number from base and novel classes are roughly equivalent. All 6970 trimmed activity videos, except the 573 distractors, are used for retrieval.

When a trimmed activity video acts as a query, the remaining videos act as the gallery.

Evaluation metrics. For video retrieval, we consider the mean average precision (mAP) both on base classes and novel classes. We also compute the harmonic mean (H) between the mAP of base classes and novel classes to evaluate the balance between base and novel class performance.

5. Results

In the experiments, we first perform a series of ablation studies and comparisons in our proposed approach. Second, we perform further analyses to gain insight into the problem and our solution. Third, we show the ability of our model to perform retrieval of video clips and moments.

5.1. Video retrieval experiments

Ablation: Visual alignment. We first investigate our visual alignment module for imbalanced activity retrieval. In Table 2a, we show the results for the baseline setting which only uses a cross-entropy loss on a linear projection of the video representations, as well as the inclusion of the module. We observe an improvement of 6.2 percent point (p.p.) for base classes and 1.9 p.p. for novel classes. Hence, for both frequent and infrequent activities, the module provides a benefit.

To understand why the visual alignment module works, we investigate the discriminative abilities of activities with and without the use of the module. Ideally, prototypes should be well separated to distill discriminative information in the embedding space. To measure the scatteredness of prototypes, we calculate the ℓ_2 distance of every pair of classes within base classes (C_{0-100}), novel classes ($C_{100-200}$), and the overall (C_{0-200}). The visual bank in the baseline is maintained by the Equation 3 without the constraint from Equation 4. We compare the scatteredness of the visual activity proposals with and without the use of the visual alignment module. The results are shown in Table 2b. The scatteredness is consistently higher with our module, making all activities more unique which in turn leads to a more discriminative retrieval.

Ablation: Semantic alignment. For the semantic alignment module, we investigate its effect using four different word embedding methods. The results are shown in Table 3a using word2vec [26], ELMo [32],GloVe [31], and fasttext [18]. For all word embedding methods, the multi-layer perceptron in semantic alignment module is kept the same except for the last layer. We find that all word embeddings provide an improvement over the setting with the baseline and our visual alignment module. For base classes, word2vec is slightly preferred, while fasttext is slightly preferred for novel classes. Overall, ELMo provides a balance between base and novel classes and we will use this word embedding for further experiments.

Having a semantic bank offers another benefit, namely

Table 2: **Ablations** on the visual alignment module.

(a) **Visual module.** Both base and novel classes benefit from the module comapred to the baseline.

method	base (mAP)	novel (mAP)	**H**
w/o visual	25.76	16.28	19.95
w/ visual	31.99	18.17	23.18

(b) **Scatteredness.** The module works because activities are distinguished well from each other.

method	base	novel	overall
baseline	0.84	0.90	0.87
+visual	**1.19**	**1.17**	**1.18**

Table 3: **Ablations** on the semantic alignment module.

(a) **Word embeddings.** Adding a semantic prior provides an improvement, regardless of the word embedding.

method	base (mAP)	novel (mAP)	**H**
baseline+visual	31.99	18.17	23.18
+word2vec [26]	33.31	18.73	23.97
+ELMo [32]	32.42	19.26	24.16
+GloVe [31]	32.59	19.28	24.23
+Fasttext [18]	32.36	19.44	24.29

(b) **Retrieval result in various activity taxonomy hierarchy**. Level-1 contains 6 super classes, level-2 contains 38 super classes. The mAP is evaluated on the overall classes.

method	level-1(6 -cls) (mAP)	level-2(38-cls) (mAP)
baseline+visual	22.41	20.35
+semantic	**23.14**	**21.76**

an enhanced retrieval performance for different levels of the activity taxonomy. We show that this is the case by utilizing the ActivityNet taxonomy [3] and evaluate the mAP for both the parent classes of the activities and the grandparent classes. The former contains 38 categories, while the latter contains 6 categories. Table 3b shows that our method is able to provide improved scores for broader activity categories, highlighting that the proposed alignment results in a semantically more coherent retrieval.

Comparison with other methods. Using our two modules, we perform a comparative evaluation to three baseline retrieval approaches. The first baseline serves as a starting point. We use the network used in this work but only pretrained on ImageNet [6] to obtain video representations by averaging their frames. Query and candidate videos are then

Table 4: **Comparison with other methods.** Our approach is preferred over both internal and external baselines, since our modules explicitly give equal importance to base and novel classes.

	base (mAP)	novel (mAP)	**H**
ImageNet [6]	9.18	13.02	10.76
Triple loss [15]	24.47	16.48	19.70
Margin loss [43]	25.84	17.36	20.76
baseline	25.76	16.28	19.95
w/ our modules	**32.42**	**19.26**	**24.16**

matched using the Euclidean distance. As result in Table 4 shows, the low scores indicate the difficulty of the task. Interestingly, the off-the-shelf baseline doesn't suffer from an imbalance performance between the base classes and novel classes. This confirms the fact that when fine-tuning, representations of videos then tend to be more discriminative towards the base classes as they are more frequent.

Table 4 also shows the consequence of imbalanced fine-tuning for two accepted approaches in retrieval, namely the triplet loss [15] and the margin loss[43] optimized on top of the same video representations as for our approach. Both approaches obtain a boost in base mAP and a smaller improvement in novel mAP. Both the sampling-based loss baselines and our baseline setup do not explicitly cater to novel classes, resulting in similar scores for the harmonic mean. Our proposed approach performs favorably compared to all baselines, both for base and for novel classes. Our harmonic mAP is respectively 13.4, 4.5, 3.4, and 4.2 percent point higher than the baselines. We conclude that our formulation is preferred for activity retrieval regardless of whether they have many or few examples to train on.

5.2. Video retrieval analyses

We perform three analyses to gain insight into the imbalanced activity retrieval problem and into our approach.

Increasing the number of samples. First, we study the effect of the number of samples per novel class during training in Figure 4. We find that even when one novel class sample is provided, our method can distill knowledge from limited provided supervision. As the number of examples for novel classes increases, the gap with the baseline also increases, highlighting that our balanced formulation also helps for many examples. We also study the effect of the number of activity videos per query during the testing phase. When using more than one query video, we average the features of all queries before retrieval. Figure 5 shows that a consistent gain can be collected when increasing the number of query videos, which shows our method benefits from having multiple videos as a query.

Robustness to data splits. As generalization over new splits is not necessarily achieved in the presence of novel classes [4, 36], we evaluate our proposed model on two
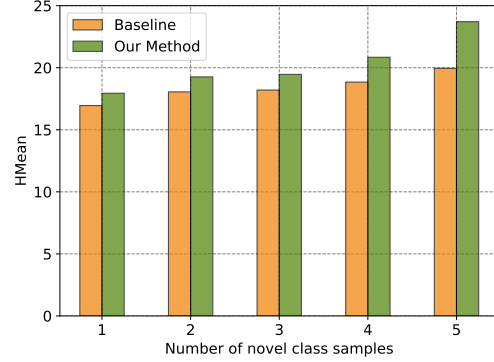

Figure 4: **Effect of the number of samples per novel class.** With our modules, the improvement gap increases with more examples per novel activity. Our balanced optimization is not only beneficial for rare activities, but also for more frequent ones.
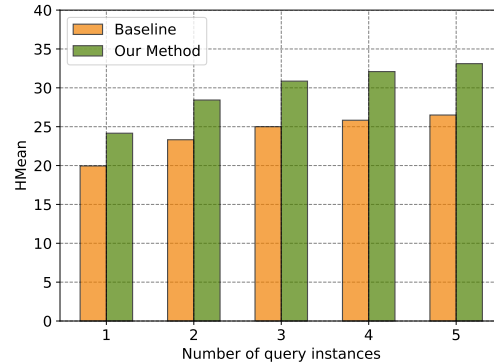

Figure 5: **Effect of the number of query videos per retrieval.** We gradually increase the number of queries from 1 to 5. Our approach is effective for both the standard and multi-shot scenario.

other data splits. Table 5 shows the results for the two following settings: (B120, N80) and (B80, N100), where B denotes the number of base classes and N the number of novel classes. The consistent improvements across these new splits indicate that our approach is not tuned to specific splits and can work whether we have many or few infrequent activities.

Qualitative analysis. Intuitively, not all activities benefit from the inclusion of our visual and semantic alignments for balancing activities. In Figure 6, we show which classes benefit and suffer the most after applying the two alignment modules. We select the 5 easiest and 5 hardest classes from the base and novel classes respectively. For the novel classes, gains are important for fine-grained activities with a salient object, such as *decorating the Christmas tree* or *carving jack-o-lanterns*. For the base classes, gains are important for sports activities. Indeed, a fine-grained understanding is required to differentiate among these activities and having both alignment modules helps to separate them. We observe that our approach suffers for multiple sports ac-

Table 5: **More dataset splitting case.** We evaluate on two difffent class label splittings: (B120, N80) and (B80, N120). Note that the original dataset splitting is (B100, N100). Our method is consistent over three different dataset splits.

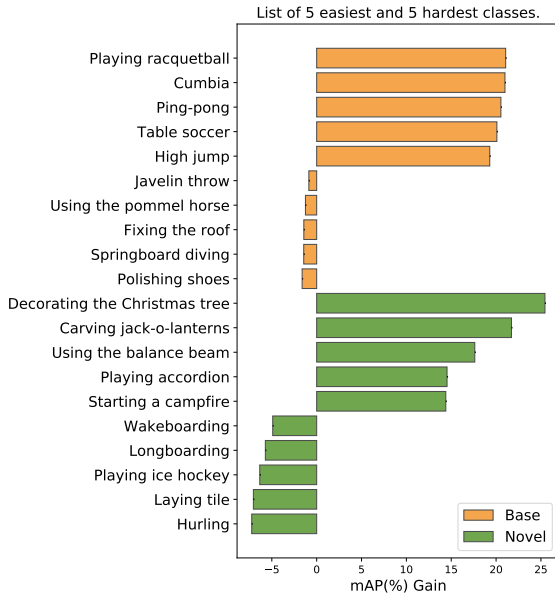|	base (mAP)	novel (mAP)	**H**
(B120, N80)			
baseline	22.72	13.41	16.86
baseline+visual	28.15	14.72	19.33
baseline+visual+semantic	**29.38**	**14.91**	**19.78**
(B80, N120)			
baseline	27.39	14.74	19.16
baseline+visual	33.14	17.14	22.59
baseline+visual+semantic	**33.85**	**17.63**	**23.18**

Figure 6: **The gain and loss analysis.** We pick out the 5 easiest and 5 hardest classes from base and novel classes respectively. The x-axis is the relative gain in percentage.

tivities with few examples, showing the direct downside of the boost for sports activities with many examples, as they will become a more likely retrieval candidate given a query.

Figure 7 also presents successful and failure cases. For the two success cases, our method can tackle various background distractors to extract essential video information. For the failure case of *cutting the grass*, our method is distracted by *e.g.* the tree in the *bungee jumping* example and by the highly-similar activity *mowing the lawn*. For the failure case of *brushing teeth*, the context information to other activities is very similar, while small key objects such as *cigarette*, *ice cream*, *shaver* are ignored by our method. Having object information could further boost the retrieval performance.

Table 6: **Clip retrieval evaluation** for clips of 4, 6, and 8 seconds. We find favourable results for all clip lengths, especially when videos are longer.

clip-duration	base (mAP)	novel (mAP)	**H**
4 seconds			
Margin loss [43]	14.06	10.10	11.76
baseline	13.38	10.33	11.66
our method	**17.62**	**12.85**	**14.86**
6 seconds			
Margin loss [43]	14.80	10.94	12.58
baseline	13.62	10.76	12.03
Our method	**18.19**	**13.36**	**15.40**
8 seconds			
Margin loss [43]	15.23	11.32	12.99
baseline	13.96	10.99	12.30
our method	**18.65**	**13.75**	**15.83**

Table 7: **Performance on Moment Retrieval.** Our method can perform favorably compared with our baseline.

	base (mAP)	novel (mAP)	**H**
Margin loss [43]	7.06	5.66	6.28
baseline	8.44	7.03	7.67
Our method	**9.14**	**7.15**	**8.02**

5.3. Clip and moment retrieval

Beyond retrieving videos, our approach is also suitable for retrieving video clips and video moments, both of which have recently gained traction. In a retrieval context, video clips denote local video segments of a fixed length [25], while video moments denote localized segments marking the duration of the activity in a whole video [13].

Clip retrieval. For clip retrieval, we set the fixed duration to 4, 6, and 8 seconds. All videos are split into fixed-length clips, where a clip is positive if its temporal during lies within the boundaries of the activity. The retrieval is performed over all individual clips.

We show clip retrieval results on the same dataset as video retrieval in Table 6. We use the most effective video retrieval baseline, the margin loss [43], as a baseline here. We find that regardless of the clip duration, our approach is preferred. As video clips become longer, the performance gap slightly increases for base and novel activities. Overall, we conclude that a generalization to video clips for retrieval is viable for our approach.

Moment retrieval. Lastly, we investigate localized video moment retrieval with our approach. We obtain temporal proposals by starting from video clips and performing a sliding window over all sets of consecutive clips exhaus-

Figure 7: **Retrieval visualization of our method.** The upper part shows two successful cases while the bottom two failure cases. We evaluate with either a query from a base or novel class. The two successful cases demonstrate our method can tackle the different background distractors and extract the essential information. The failing *cutting the grass* is distracted by green grass in distractor video, green tree in *bungee junmping*, context information in *mowing the lawn*. The failure case in *Brushing teeth* contains a similar context information, while still fails. The object recognition of *cigarette*, *ice cream*, or *shaves* would be helpful for the retrieval task.

Figure 8: **VR-ActivityNet statistics about the max moment length (M) and clip length (N).** We exhaustively list all possible moment temporal proposals in all videos of our dataset. A hit occurs when the tIoU between a proposal with class C and one ground truth proposal with class C is larger than 0.5. The best combination is N=5, M=26, which is our default setting in moment retrieval task.

tively. We observe that such a proposal setup readily obtains proposals with high recall, as shown in Figure 8. For retrieval, we score each proposal in each video and rank them by similarity score. Table 7 shows that our method can also generalize to video moment retrieval with imbalanced activities. The improvements are more marginal compared

to video and clip retrieval, since moment retrieval entails a more difficult task due to the additional temporal localization.

6. Conclusion

In this work we propose a new task about video retrieval by activity in the wild, and emphasize the importance of dealing with imbalanced data when retrieving activities from a video query. We introduce an embedding network that learns to balance frequent base activities and infrequent novel activities. The network contains two novel modules. A visual alignment module matches input videos with visual prototype representations of activities. A semantic alignment module on the other hand matches videos with word embedding representations of activities. Visual and semantic activity representations are of the same length, regardless of the number of examples each activity has. As a result, we arrive at an activity retrieval that better balances both types of activities. We show this result empirically by proposing a new imbalanced activity retrieval dataset with a revised data splits. Experiments highlight the effectiveness of our approach, as well as a series of ablations and analyses to gain insight into the problem. Lastly, we show how our approach generalizes to video clip and moment retrieval from video queries in imbalanced settings.

References

[1] Evgeniy Bart and Shimon Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In *CVPR*, 2005. 2

[2] Samuel Rota Bulò, Gerhard Neuhold, and Peter Kontschieder. Loss max-pooling for semantic image segmentation. *CoRR*, 2017. 1, 2

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 2, 4, 5, 12

[4] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICCV*, 2019. 6

[5] Arridhana Ciptadi, Matthew S Goodwin, and James M Rehg. Movement pattern histogram for action recognition and retrieval. In *ECCV*, 2014. 1, 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 4, 5, 6

[7] Matthijs Douze, Jérôme Revaud, Jakob Verbeek, Hervé Jégou, and Cordelia Schmid. Circulant temporal encoding for video retrieval and temporal alignment. *IJCV*, 2016. 1

[8] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. *arXiv*, 2019. 1, 2

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *ICML*, 2017. 2

[10] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 2

[11] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *ICCV*, 2017. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[13] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 1, 2, 7

[14] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. *arXiv*, 2018. 2

[15] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, 2015. 6

[16] Tao Hu, Pascal Mettes, Jia-Hong Huang, and Cees GM Snoek. Silco: Show a few images, localize the common object. In *ICCV*, 2019. 3

[17] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv*, 2017. 4

[18] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv*, 2016. 5

[19] In-Ho Kang and GilChang Kim. Query type classification for web document retrieval. In *ACM Conference on Research and Development in Information Retrieval*, 2003. 1

[20] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *Conference on Medical Imaging with Deep Learning*, 2019. 1, 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2016. 4

[22] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019. 1, 2

[23] Tiange Luo, Aoxue Li, Tao Xiang, Weiran Huang, and Liwei Wang. Few-shot learning with global class representations. *arXiv*, 2019. 1, 2

[24] Pascal Mettes, Elise van der Pol, and Cees G. M. Snoek. Hyperspherical prototype networks. *arXiv*, 2019. 1

[25] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 2, 7

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013. 4, 5

[27] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *CVPR*, 2019. 2

[28] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *arXiv*, 2019. 1, 2

[29] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. In *ECCV*, 2016. 2

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NeurIPS*, 2017. 4

[31] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 5

[32] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv*, 2018. 4, 5

[33] Jie Qin, Li Liu, Mengyang Yu, Yunhong Wang, and Ling Shao. Fast action retrieval from videos via feature disaggregation. *Computer Vision and Image Understanding*, 2017. 1

[34] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv*, 2018. 2

[35] Joan Serrà, Dídac Surís, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. *arXiv*, 2018. 2

[36] Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv*, 2018. 6

[37] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 1

[38] Cees GM Snoek, Marcel Worring, et al. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2009. 1

[39] Jingkuan Song, Hanwang Zhang, Xiangpeng Li, Lianli Gao, Meng Wang, and Richang Hong. Self-supervised video hashing with hierarchical binary auto-encoder. *IEEE Transactions on Image Processing*, 2018. 1, 2

[40] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv*, 2016. 2

[41] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 3

[42] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, 2019. 3

[43] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *ICCV*, 2017. 6, 7, 12

[44] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 2017. 11

[45] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, 2019. 2

[46] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *ECCV*, 2018. 1

[47] Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. Capturing long-tail distributions of object subcategories. In *CVPR*, 2014. 1, 2

7. Supplementary File

Contents

    A. Extra experimental results

    A.1 Duration analysis. We analyze the performance with respect to the duration of the query videos in Figure 9. We can observe that the longer the query video is, the more performance gain our method can collect, which indicates the efficacy of our method.

    A.2 mAP curve. mAP is an important metric to evaluate the balance between the precision and recall of retrieval task. All retrieval results per query are considered in mAP. We depict the mAP in Figure 10. The performance gap between base and novel classes is still relatively large, more improvement such as object recognition ability can be further embedded to boost the performance.

    A.3 Confusion matrix. As the original number of activity class labels is 200, we randomly select 20 classes to illustrate the confusion matrix in Figure 11. The top-100 retrieval will be all seen as a hit in the generation of confusion matrix. We observe that *canoeing* performs best because the context information is not so complex, while *painting* presents the worst result among those classes. This probably comes from the context information in *painting*, which is more variable. Also, we can obverse that *arm wrestling* is seriously mistaken as *playing flauta* as both activities show arms. These observations also align with the phenomenon in the visualization results. The ability of object recognition could further boost the performance of this task.

    A.4 Loss curve. We show our loss decreasing curve in Figure 12. Three different kinds of loss items are depicted. A globally stable decreasing trend can be observed.

    A.5 Dataset label splits. We summarize the dataset label split in Table 8. Those labels are randomly splited into three subsets.

    B. Implementation details

    B.1 Evaluation Metric Calculation. We use mAP as the main metric, for a better balance of performance between base and novel classes, similar to [44], we consider the harmonic mean between the base and novel mAP as our evaluation metrics. The calculation of mAP of base classes and novel classes is:

$$mAP = \frac{1}{M} \sum_{i=1}^{K} \sum_{j=1}^{Q_c} AP_{i,j} \tag{7}$$

    where K is the total class number, Q_k is the query number from class k, $AP_{k,j}$ denotes the AP of class from k, query from j. M is the total query video number from base or novel classes.
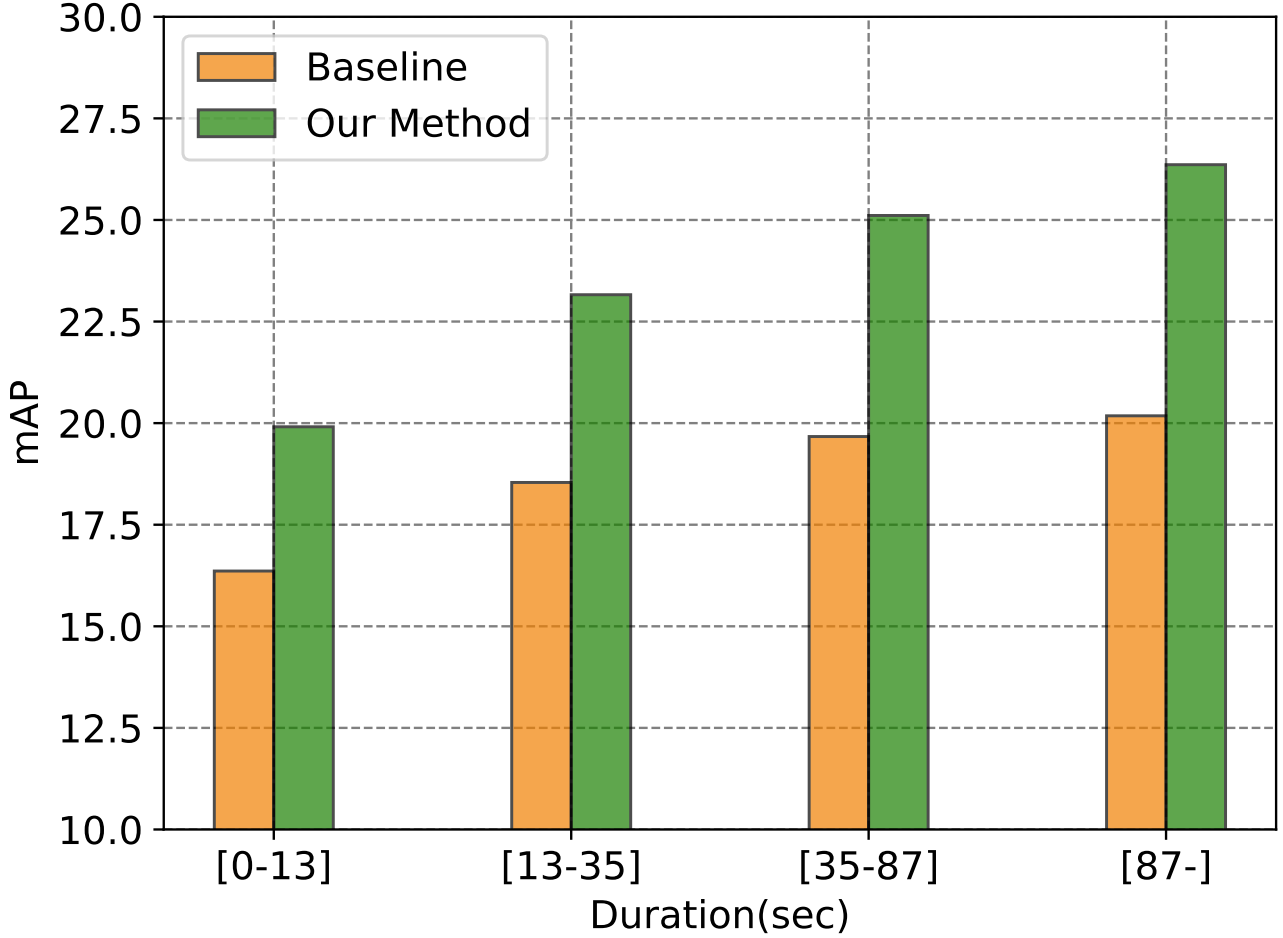
Figure 9: **Retrieval performance with respect to the query video duration (sec).** Our method is able to collect consistent gain as duration increases, which demonstrates a better feature fusion efficacy can be achieved by our method.

B.2 Baseline details. Triplet loss. We randomly sample 2 classes to formulate the anchor, positive, negative class. Based on these classes, we further randomly sample three videos to construct a triplet for training. **Margin loss [43].** The margin parameter is set to 0.2.

B.3 FC Layer in Semantic Alignment module. Let FC(M) denote a fully connected layer with M units. ReLU denotes the activation function and S means the dimension of the specific word embedding method. The FC Layer in Semantic Alignment Module can be formulated as: FC(512)-ReLU-FC(640)-ReLU-FC(768)-ReLU-FC(896)-ReLU-FC(S).

B.4 ActivityNet hierarchy label preprocessing. Our hierarchy label is slightly different from the official annotation of ActivityNet [3], we tidy up the hierarchy structure based on the official version to formulate a 6 grandfather activity categories, 38 father activity categories, and 200 activity classes. Our hierarchy structure will be released public upon acceptance.
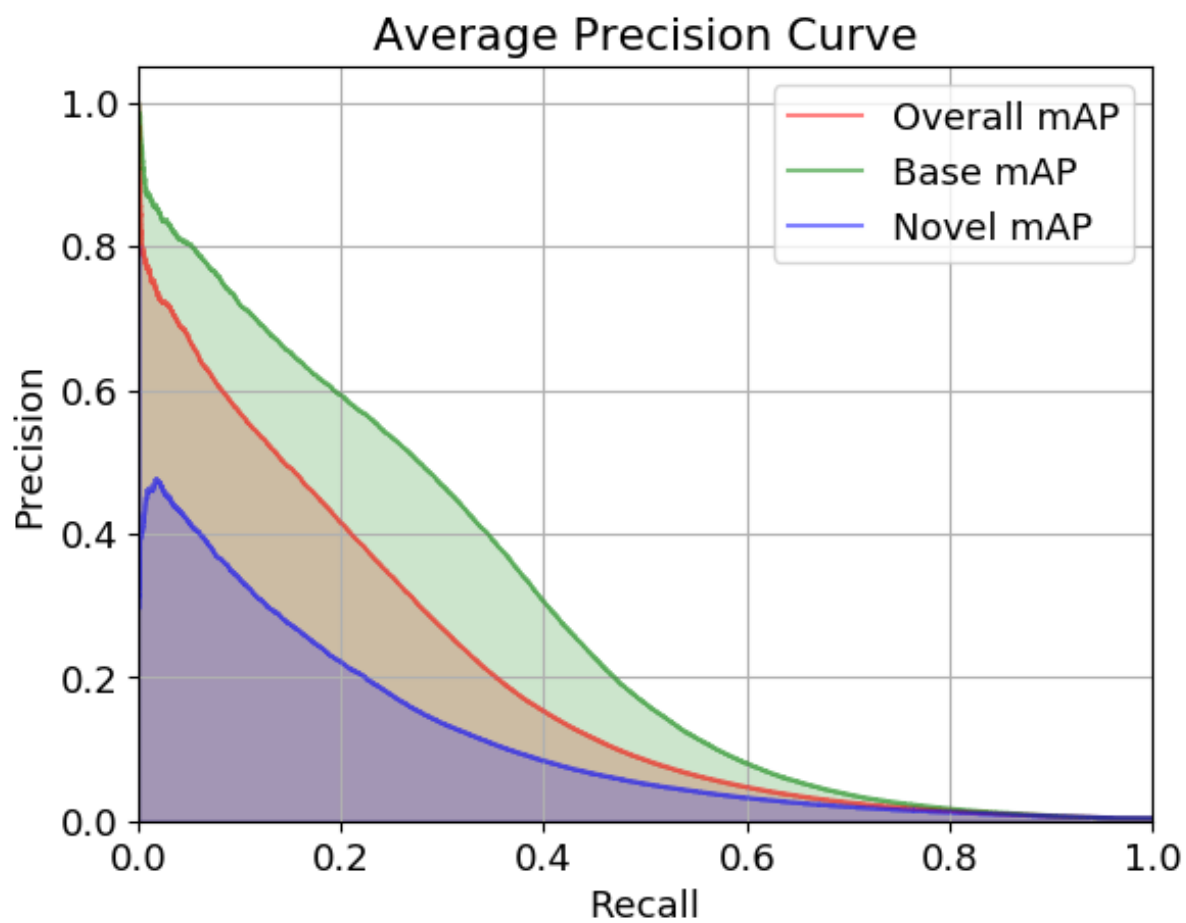
Figure 10: **mAP curve of our best result,** which shows the base, novel and overall classes.
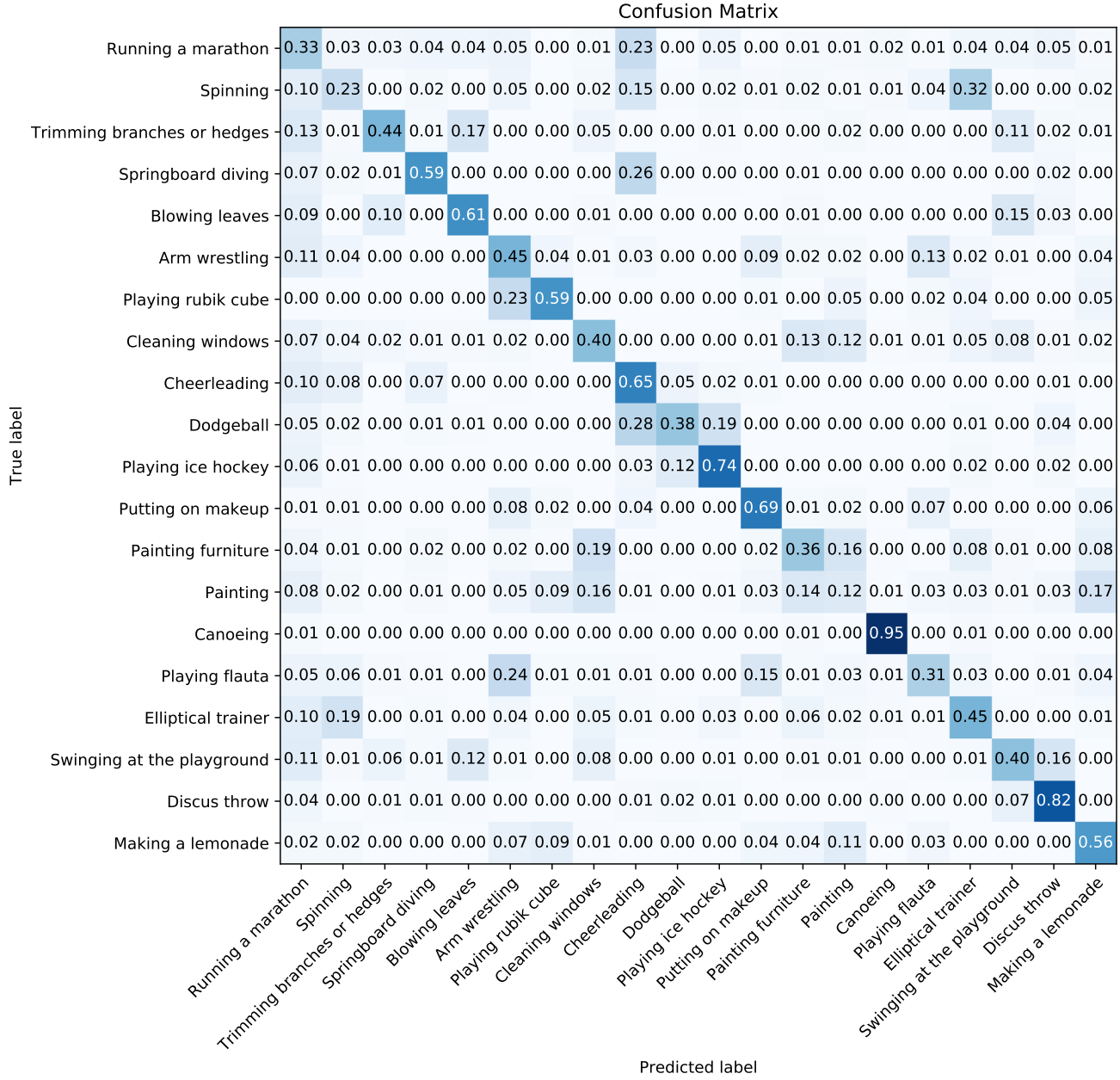
Figure 11: **Confusion matrix** demonstrate that our method can deal with most of the cases well. Vertical axis means ground truth, horizontal axis denotes prediction. The 20 classes are random selected from the 200 classes. The top-100 will seen as a hit in the generation of confusion matrix.
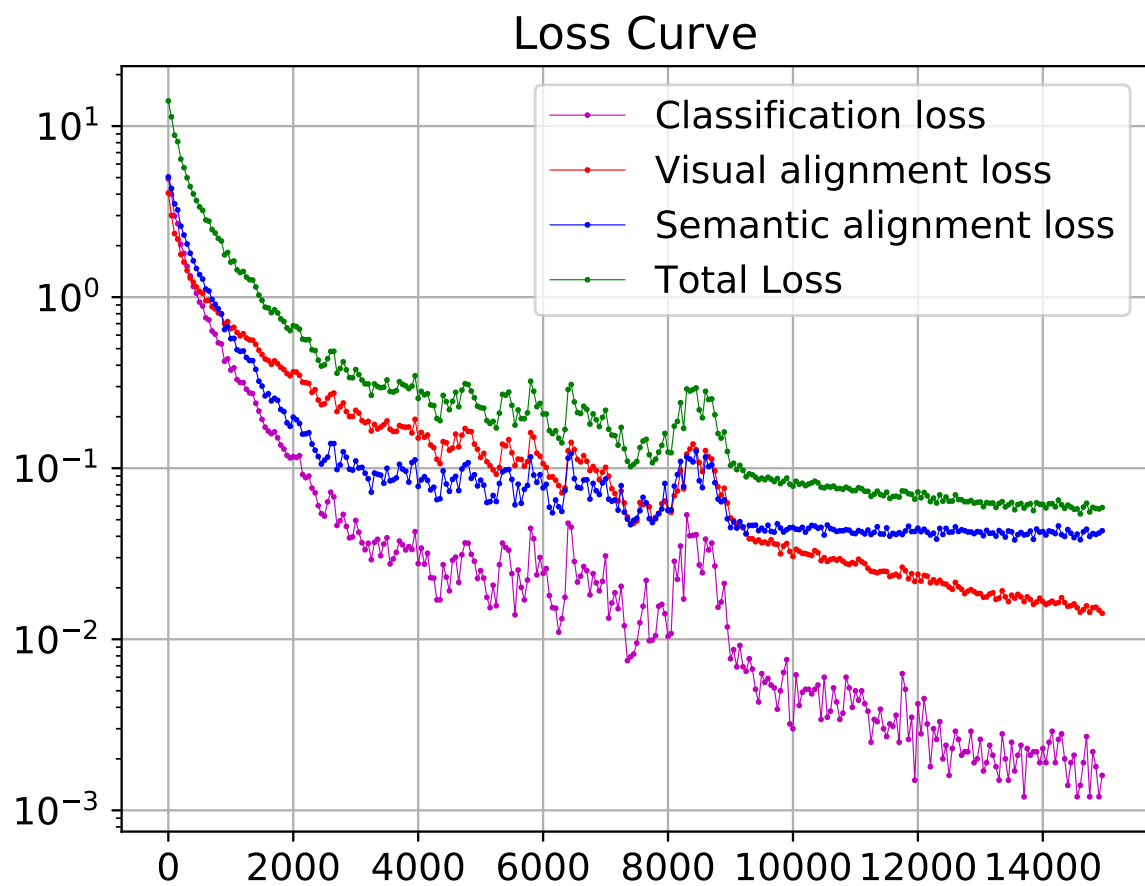
Figure 12: **Loss Curve.** We show four items: classification loss, visual alignment loss, semantic alignment loss and total loss. Log scale in Y-axis.

Table 8: **Dataset labels.** The validation set is trying to evaluate the balance performance between the C_{0-100} and $C_{100-120}$, similarly, the testing set is designed to evaluate the balance performance between the C_{0-100} and $C_{120-200}$.

subset	labels
Base: C_{0-100}	Cricket, Cleaning windows, Cutting the grass, Roof shingle removal, Cheerleading, Skiing, Using parallel bars, Putting in contact lenses, Doing step aerobics, Bathing dog, Springboard diving, Camel ride, Horseback riding, Installing carpet, Washing dishes, Grooming dog, Getting a piercing, Rafting, Using the monkey bar, Rock-paper-scissors, Ping-pong, Washing hands, Arm wrestling, Putting on shoes, Snatch, Doing fencing, Chopping wood, Fixing the roof, Doing kickboxing, River tubing, Belly dance, Spinning, Waxing skis, Layup drill in basketball, Ice fishing, Zumba, Slacklining, Sumo, Kayaking, Playing congas, Doing nails, Playing harmonica, High jump, Playing racquetball, Table soccer, Preparing salad, Running a marathon, Doing motocross, Smoking a cigarette, Hitting a pinata, Polishing forniture, Tennis serve with ball bouncing, Hula hoop, Disc dog, Hopscotch, Smoking hookah, Bullfighting, Hammer throw, Javelin throw, Cumbia, Paintball, Dodgeball, Baking cookies, Playing saxophone, Futsal, Making a sandwich, Riding bumper cars, Rope skipping, Removing ice from car, Long jump, Welding, Clean and jerk, Curling, Using the pommel horse, Mooping floor, Drinking coffee, Playing rubik cube, Playing water polo, Wrapping presents, Skateboarding, Removing curlers, Painting fence, Preparing pasta, Having an ice cream, Doing crunches, Grooming horse, Trimming branches or hedges, Shoveling snow, Kite flying, Playing violin, Blow-drying hair, Shuffleboard, Playing bagpipes, Playing drums, Polishing shoes, Playing piano, Playing polo, Blowing leaves, Archery, Brushing hair
Novel: $C_{100-120}$	Hurling, Laying tile, Rock climbing, Doing a powerbomb, Throwing darts, Using the balance beam, Playing field hockey, Elliptical trainer, Raking leaves, Painting furniture, Capoeira, Playing ice hockey, Snowboarding, Hand car wash, Baton twirling, Changing car wheel, Playing lacrosse, Playing beach volleyball, Tai chi, Knitting
Novel: $C_{120-200}$	Shaving, Gargling mouthwash, Getting a haircut, Vacuuming floor, Making a cake, Triple jump, Discus throw, Painting, Swinging at the playground, Ballet, Using the rowing machine, Shaving legs, Washing face, Braiding hair, Kneeling, Peeling potatoes, Snow tubing, Calf roping, Playing badminton, Tango, Bungee jumping, Playing accordion, Fun sliding down, Tumbling, Making a lemonade, Volleyball, Spread mulch, Rollerblading, Fixing bicycle, Drum corps, Pole vault, Cleaning shoes, Beach soccer, Windsurfing, Powerbocking, Croquet, Ironing clothes, Starting a campfire, Walking the dog, Getting a tattoo, Plataform diving, Sharpening knives, Tug of war, Hanging wallpaper, Surfing, Mowing the lawn, Assembling bicycle, Wakeboarding, Cleaning sink, Doing karate, Building sandcastles, Carving jack-o-lanterns, Waterskiing, Applying sunscreen, Brushing teeth, Playing squash, Playing pool, Scuba diving, BMX, Swimming, Decorating the Christmas tree, Longboarding, Sailing, Using uneven bars, Playing guitarra, Mixing drinks, Playing flauta, Drinking beer, Hand washing clothes, Playing kickball, Playing ten pins, Playing blackjack, Canoeing, Breakdancing, Shot put, Plastering, Beer pong, Clipping cat claws, Putting on makeup, Making an omelette